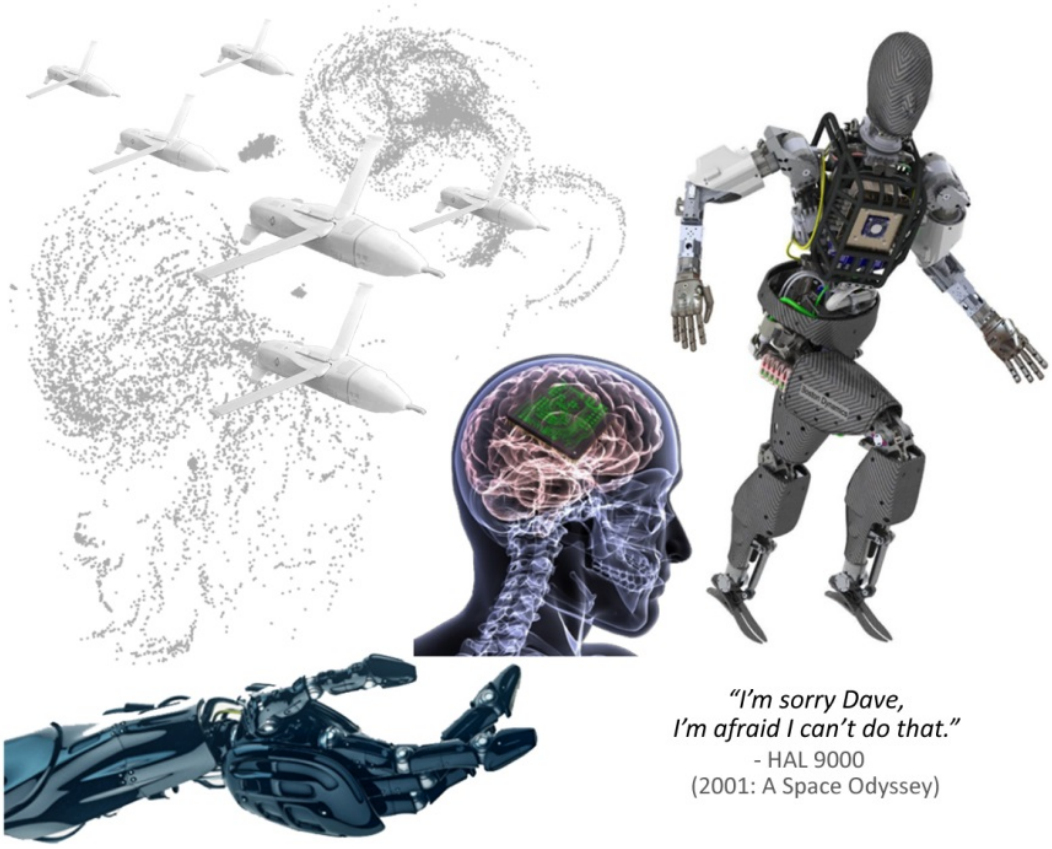
AI, Robots, and Swarms

Issues, Questions, and Recommended Studies

Andrew Ilachinski

January 2017



CNA
ANALYSIS & SOLUTIONS

This document contains the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the sponsor.

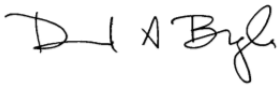
Distribution

Approved for Public Release; Distribution Unlimited. Specific authority: N00014-11-D-0323.

Copies of this document can be obtained through the Defense Technical Information Center at www.dtic.mil or contact CNA Document Control and Distribution Section at 703-824-2123.

Photography Credits: http://www.darpa.mil/DDM_Gallery/Small_Gremlins_Web.jpg; http://4810-presscdn-0-38.pagely.netdna-cdn.com/wp-content/uploads/2015/01/Robotics.jpg; http://i.kinja-img.com/gawker-edia/image/upload/18kxb5jw3e01ujpg.jpg

Approved by: **January 2017**



Dr. David A. Broyles
Special Activities and Innovation
Operations Evaluation Group

Abstract

The military is on the cusp of a major technological revolution, in which warfare is conducted by unmanned and increasingly autonomous weapon systems. However, unlike the last "sea change," during the Cold War, when advanced technologies were developed primarily by the Department of Defense (DoD), the key technology enablers today are being developed mostly in the commercial world. This study looks at the state-of-the-art of AI, machine-learning, and robot technologies, and their potential future military implications for autonomous (and semi-autonomous) weapon systems. While no one can predict how AI will evolve or predict its impact on the development of military autonomous systems, it is possible to anticipate many of the conceptual, technical, and operational challenges that DoD will face as it increasingly turns to AI-based technologies. This study examines key issues, identifies analysis gaps, and provides a roadmap of opportunities and challenges. It concludes with a list of recommended future studies.

This page intentionally left blank.

Executive Summary / White Paper

A notable number of groundbreaking artificial intelligence (AI)-related technology announcements and/or demonstrations took place in 2016:[1]

1. AI defeated the reigning world champion in the game of Go, a game that is so much more "complex" than chess that, prior to this event, most AI experts believed that *it could not be done for another 15-20 years*.[2]
2. AI learned—*on its own*—where to find the information it needs to accomplish a specific task.[3]
3. AI predicted the immediate future (by generating a short video clip) by *examining a single photograph* (and is also able to predict the future from studying video frames).[4]
4. AI automatically inferred the rules that govern the behavior of individual robots within a robotic swarm *simply by watching*.[5]
5. AI learned to navigate the London Underground *by itself* (by consulting its own acquired memories and experiences, much like a human brain).[6]
6. AI speech recognition reached human parity in conversational speech.[7]

[1] Most of the innovations on this list are described in the *Artificial Intelligence* section of the main narrative of this report (pp. 44-71). A few others also appear in the appendix.

[2] C. Koch, "How the Computer Beat the Go Master," *Scientific American*, 19 March 2016.

[3] K. Narasimhan et al., "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning," presented at EMNLP 2016, https://arxiv.org/abs/1603.07954.

[4] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating Videos with Scene Dynamics," presented at the 29th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: http://web.mit.edu/vondrick/tinyvideo/paper.pdf.

[5] W. Li, M. Gauci, and R. Gross, "Turing learning: a metric-free approach to inferring behavior and its application to swarms," *Swarm Intelligence* 10, no. 3, September 2016: http://link.springer.com/article/10.1007%2Fs11721-016-0126-1.

[6] E. Gibney, "Google's AI reasons its way around the London Underground," *Nature*, Oct 2016.

[7] X. Xiong et al., "Achieving Human Parity in Conversational Speech Recognition," *arXiv*, 2016: https://arxiv.org/abs/1610.05256.

7. An AI communication system *invented its own encryption scheme*, without being taught specific cryptographic algorithms (and without revealing to researchers how its method works).[8]
8. An AI translation algorithm invented its own "interlingua" language to more effectively translate between any two languages (*without being taught to do so by humans*).[9]
9. An AI system *interacted with its environment* (via virtual actuators) to learn and solve problems in the same way that a human child does.[10]
10. An AI-based medical diagnosis system at the Houston Methodist Research Institute in Texas achieved 99% accuracy in reviewing millions of mammograms (at a rate 30× faster than humans).[11]

These and other recent similar breakthroughs (e.g., IBM's *Watson's* defeat of the two highest ranked *Jeopardy!* players of all time in 2011),[12] are notable for several reasons. First, they collectively provide evidence that we, as a species, have already crossed over into an era in which seeing AI outperform humans—at least for specific tasks—is *almost* routine (perhaps in the same way that landing on the moon was "almost" routine after the first few Apollo missions).[13] Second, they offer a glimpse of how *different* AI is from human intelligence, and how inaccessible its "thinking" is to outside probes. And third, they demonstrate the power of AI to *surprise* us (including AI system developers, who nowadays are closer in spirit to "data collectors" and "trainers" than to traditional programmers)—i.e., AI, at its core, is fundamentally *unpredictable*. In the second game of the Go match between the AI that defeated Lee SeDol (an 18-time world champion in Go), the AI made a move so surprising that

[8] M. Abadi and D. Andersen, "Learning to Protect Communications with Adversarial Neural Cryptography," arXiv:1610.06918v1: https://arxiv.org/abs/1610.06918.

[9] Q. Le and M. Schuster, "A Neural Network for Machine Translation, at Production Scale," Google Research Blog, 27 Sep 2016: https://research.googleblog.com/2016/09/a-neural-network-for-machine.html.

[10] M. Denil, P. Agrawal, T. Kulkarni, et al., "Learning to perform physics experiments via deep reinforcement learning," under review as a conference paper to ICLR 2017: https://arxiv.org/pdf/1611.01843v1.pdf.

[11] T. Patel et al., "Correlating mammographic and pathologic findings in clinical decision support using NLP and data mining methods," *Cancer* 123, 1 Jan 2017.

[12] S. Baker, *Final Jeopardy: Man vs. Machine and the Quest to Know Everything*, Houghton Mifflin Harcourt, 2011.

[13] Unlike the Apollo program, however, AI is here to stay: *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence*, Report of the 2015 Study Panel, Stanford University, Sep 2016.

SeDol had to leave the room for 15 minutes to recover his composure: "It's not a human move. I've never seen a human play this move. So beautiful."[14]

The breakthroughs listed above are also notable for a fourth reason—a more subtle one, but the one that directly inspired this study. Namely, they portend a set of deep conceptual and technical challenges that the Department of Defense (DoD) must face, now and in the foreseeable future, as it embraces *AI*–, *robot*–, and *swarm*–related technologies to enhance (and weaponize) its fleet of unmanned systems with higher levels of autonomy. The subtlety lies in unraveling the true meaning of the deceptively "obvious" word, *autonomy;* indeed, as of this writing, there is no universally accepted definition.

Autonomous weapons—colloquially speaking—have been used since World War II (e.g., the German *Wren* torpedo's passive acoustic homing seeker effectively made it the world's first autonomously guided munition).[15] Human-supervised automated defensive systems have existed for decades, and aerial drones were first used more than 20 years ago (i.e., the RQ-1 Predator was used as an intelligence, surveillance, and reconnaissance platform in former Yugoslavia).[16] But it was only after the September 11, 2001, terrorist attacks that the military's burgeoning interest in, and increasing reliance on, unmanned vehicles started in earnest. In just 10 years, DoD's inventory of unmanned aircraft grew from 163, in 2003, to close to 11,000, in 2013 (and, in 2013, accounted for 40% of *all* aircraft).[17] And the United States is far from being alone in its interest in drones: by one recent tally, at least 30 countries have large military drones, and the *weaponized* drone club has recently grown to 11 nations, including the United States.[18]

DoD procured most of its medium-sized and larger unmanned aerial vehicles (UAVs), the MQ-1/8/9s and RQ-4/11s, for the counterinsurgency campaigns in Iraq and Afghanistan, where the airspace was largely uncontested. Now the United States is withdrawing from those campaigns and the military is shifting its strategic focus to less permissive operating environments (i.e., the Asia-Pacific region) and to adversaries with modern air defense systems. Thus, there is a growing emphasis on developing new, more *autonomous*, systems that are better equipped to survive in more contested airspaces.

[14] C. Metz, "The Sadness and Beauty of Watching Google's AI play Go," *Wired*, 11 March, 2016.

[15] J. Campbell, *Naval Weapons of World War Two*, Naval Institute Press, 2002.

[16] P. Springer, *Military Robots and Drones: A Reference Handbook*, ABC-CLIO, 2013.

[17] *Unmanned Systems Integrated Roadmap: FY2013-2038*, U.S. Department of Defense, 2013.

[18] *World of Drones: Military*, International Security Data Site, New America Foundation: http://securitydata.newamerica.net/world-drones.html.

Fundamentally, an autonomous system is a system that can independently compose and select among alternative courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the local, dynamic context. Unlike automated systems, autonomous systems must be able to respond to situations that are not pre-programmed or anticipated prior to their deployment. In short, autonomous systems are inherently, and irreducibly, *artificially intelligent robots*. In the remaining pages of this summary, we explicate the analytical implications of this assertion (leaving details and supporting evidence to the main narrative).

To start, if and when autonomous systems, in the sense just described, finally arrive, they will offer a variety of obvious advantages to the warfighter. For example, they will eliminate the risk of injury and/or death to the human operator; offer freedom from human limits on workload, fatigue, and stress; and be able to assimilate high-volume data and make "decisions" based on time scales that far exceed human ability. If robotic swarms are added into the mix, entirely new mission spaces potentially open up as well—e.g., wide-area, long-persistence, surveillance; networked, adaptive electronic jamming; and coordinated attack. There are also numerous advantages to using swarms rather than individual robots, including: *efficiency* (if tasks can be decomposed and performed in parallel), *distributed action* (multiple simultaneous cooperative actions can be performed in different places at the same time), and *fault tolerance* (the failure of a single robot within a group does not necessarily imply that a given task cannot be accomplished).

However, the design and development of autonomous systems also entails significant conceptual and technical challenges, including:

- *"Devil is in the details" research hurdles:* Developers of autonomous systems must confront many of the same fundamental problems that the academic and commercial AI and robotic research communities have struggled for decades to "solve." To survive and successfully perform missions, autonomous systems must be able to sense, perceive, detect, identify, classify, plan for, decide on, and respond to a diverse set of threats in complex and uncertain environments. While aspects of all these "problems" have been solved to varying degrees, there is, as yet, no system that fully encompasses all of these features.

- *Complex and uncertain environments:* Autonomous systems must be able to operate in complex—possibly, a priori unknown—environments that possess a large number of potential states that cannot all be pre-specified or be exhaustively examined or tested. Systems must be able to assimilate, respond to, and adapt to dynamic conditions that were not considered during their design. This "scaling" problem—i.e., being able to design systems that are developed and tested in static and structured environments, and then have

them perform as required in dynamic and unstructured environments—is highly nontrivial.

- *Emergent behavior:* For an autonomous system to be able to adapt to changing environmental conditions, it must have a built-in capacity to learn, and to do so without human supervision. It may be difficult to predict, and be able to account for *a priori* unanticipated, emergent behavior (a virtual certainty in sufficiently "complex" systems-of-systems dynamical systems).

- *Human-machine interactions/I:* The operational effectiveness of autonomous systems will depend on the dynamic interplay between the human operator and the machine(s) in a given environment, and on how the system responds, in real time, to changing operational objectives, in concert with the human's own adaptation to dynamic contexts. The innate unpredictability of the human component in human-machine collaborative performance only exacerbates the other challenges identified on this list.

- *Human-machine interactions/II:* The interface between human operators and autonomous systems will likely include a diverse space of tools that include visual, aural, and tactile components. In all cases, there is the challenge of translating human goals into computer instructions (e.g., "solving" a long-standing "AI problem" of natural language processing), as well as that of depicting the machine's "decision space" in a form that is understandable by the human operator (e.g., allowing the operator to answer the question, "Why did the system choose to take action X?").

- *Control:* As autonomous systems increase in complexity, we can expect a commensurate decrease in our ability to both predict and control such systems—i.e., the "spectre of complacency in complexity." As evidenced by the general nature of recent AI breakthroughs, there is a fundamental tradeoff: either the AI can achieve a given performance level (e.g., it can play the game Go as well as, or better than, a human), or humans can be able to understand how its performance is being achieved).

Apart from these innately technical challenges to developing autonomous systems, there are a set of concomitant acquisition challenges, the origin of which is a recent shift in DoD's innovation-related procurement practices. While the U.S. government has always played an important role in fostering AI research (e.g., ARPA, DARPA, NSF, ONR), most key innovations in AI, robotics, and autonomy are now being driven by the *commercial sector*,[19] and at a pace that DoD's relatively plodding stove-piped

[19] The development of most of the UAVs used in Iraq and Afghanistan was driven not by DoD requirements, but rather by commercial research and development. Ref: "Microsoft, Google,

acquisition process is ill equipped to accommodate: it takes 91 months (7.6 years), on average, from the start of an analysis of alternatives (AoA) study to initial operational capability (IOC).[20] Even information technology programs—under whose rubric most AI-derived acquisitions naturally fall—have averaged 81 months. By way of comparison, note that within roughly this same interval of time, the commercial AI research community has gone from just *experimenting* with (prototypes of dedicated hardware-assisted) deep learning techniques,[21] to beating the world champion in Go (along with achieving many other major breakthroughs).

Of course, DoD acquisition challenges, particularly for weapons systems that include a heavy coupling between hardware and software, have been known for decades.[22] However, despite numerous attempts by various stakeholders to address these challenges, the generic acquisition process (at least on the traditional institutional level) remains effectively unchanged. Whatever progress has been made in recent years derives more from *workarounds* instituted by DoD to facilitate "rapid acquisition" of systems,[23] than from wholesale changes applied to stove-piped processes of the acquisition process itself. Some recent progress has been made—e.g., the 2009/2011 National Defense Authorization Acts (NDAA/Sec 804), mandated a new IT acquisition process, which, in turn led to multiple Defense Science Board (DSB) Task Force (TF) studies of the acquisition process. Yet, a notable absence in any of these DSB/TF studies is any explicit mention of autonomy.

Complicating the issue still further is a basic dichotomy between DoD's existing directive on autonomy (DoD Directive 3000.09, issued Nov 2012) and current Test and Evaluation (T&E) and Verification and Validation (V&V) practices. Specifically,

---

Facebook and more are investing in artificial intelligence: What is their plan and who are the other key players?" *TechWorld*, September 29, 2016.

[20] *Policies and Procedures for the Acquisition of Information Technology*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, March 2009.

[21] The first graphics-processor-based unsupervised deep-learning techniques were introduced in 2009: R. Raina, A. Madhavan, and A. Ng, "Large-scale deep unsupervised learning using graphics processors," *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009.

[22] J. Merritt and P. Sprey, "Negative marginal returns in weapons acquisition," *in American Defense Policy*, Third Edition, edited by R. Head and E. Roppe, John Hopkins Univ. Press, 1973.

[23] Examples include: the U.S. Air Force Rapid Capabilities Office, the U.S. Army's Asymmetric Warfare Group and Rapid Capabilities Office, DoD's Strategic Capabilities Office, and, most recently, SecDef Ashton Carter's Defense Innovation Unit Experimental (DIUx). Ref: B. Fitzgerald, A. Sander, J. Parziale, *Future Foundry: A New Strategic Approach to Military-Technical Advantage*, Center for a New American Security, 2016.

Directive 3000.09 requires that weapons systems (italics added by author of this report):[24]

- Go through rigorous hardware and software T&E/V&V, "including analysis of *unanticipated emergent behavior* resulting from the effects of complex operational environments on autonomous or semiautonomous systems."
- "Function as anticipated in realistic operational environments against *adaptive adversaries.*"
- "Are sufficiently robust to minimize failures that could lead to *unintended engagements.*"

Directive 3000.09 further requires that T&E/V&V must "assess system performance, capability, reliability, effectiveness, and suitability under realistic conditions, including possible adversary actions, consistent with the *potential consequences of an unintended engagement or loss of control of the system.*"

Yet, existing T&E/V&V practices do not make accommodations for any of the italicized parts of these quoted requirements. Among the many reasons why autonomous systems are particularly difficult to test and validate are: (1) *complexity of the state-space* (it is impossible to conduct an exhaustive search of the vast space of possible system "states" for autonomous systems); (2) *complexity of the physical environment* (the behavior of an autonomous system cannot be specified—much less tested and certified—in situ, but must be tested in concert with interaction with a dynamic environment, rendering the space of system inputs/outputs and environmental variables combinatorically intractable); (3) *unpredictability* (to the extent that autonomous systems are inherently complex adaptive systems, novel or unexpected behavior can be expected to arise naturally and unpredictably in certain dynamic situations; existing T&E/V&V practices do not have the requisite fidelity to deal with emergent behavior); and (4) *human operator trust in the machine* (existing T&E/VV&A practice is limited to testing systems in closed, scripted environments, since "trust" is not an innate trait of a system).

Trust also entails grappling with the issue of *experience* and/or *learning:* to be more effective, autonomous systems may be endowed with the ability to accrue information and learn from experience. But such a capability cannot be certified monolithically, during one "check the box" period of time. Rather, it requires periodic retesting and recertification, the periodicity of which is necessarily a function of the system's history and mission experience. Existing T&E/V&V practices are wholly inadequate to address these issues.

[24] Enclosures 2 and 3 of DoD Directive 3000.09 (*Autonomy in Weapon Systems*, Nov 2012) address T&E and V&V issues, and generally review guidelines, respectively.

Defining autonomy

"Autonomy" applies to a vastly greater range of processes than those that pertain to unmanned vehicles—as physical entities—alone, including the myriad factors needed to describe human-machine interactions. It represents a range of *context-dependent capabilities* that may appear at different scales, and in varying degrees of sophistication, that collectively enable the coupled human-machine system to perform specific tasks. Autonomy—by itself—does not reductively "fix" any existing problems; rather, it redefines, extends, and potentially opens up entirely new mission spaces. And its value can only be assessed in the context of specific mission requirements, the operating environment, and its coupling with human operators.

A major impediment to the development of autonomous weapon systems is the current lack of a common language by which AI, robot, and other technology experts, systems developers, and program managers can communicate (in a manner consistent with autonomy's multi-dimensional, context-dependent nature). There is not an even a single definition of the *word* "autonomy," much less a universally agreed upon taxonomy that might be used as basis for forming a common language. Some taxonomies emphasize the details related to a system's output functions (i.e., to its decision capability), while others focus on making detailed distinctions between input functions, such as how a system acquires information and how it formulates options. And, while sliding scales have been used to delineate between levels of "human control" that a given system might require (e.g., the "autonomy" of a system may be ranked from, say, 0, meaning that it is under complete control, to 10, meaning it is fully autonomous, albeit, typically, without the term "fully" being well defined), the practical utility of these kinds of taxonomies is limited because they ignore critically important contextual factors. For this reason, a recent U.S. Defense Science Board report recommended doing away with defining levels of autonomy altogether and replacing such taxonomies with a comprehensive conceptual framework. However, to date, despite a handful of ongoing attempts, no useable framework yet exists.

Ethical concerns

The emerging use of autonomous weapons—and the spectre (if not yet the reality) of *lethal* autonomous weapon systems (LAWS), that can select and engage targets on their own[25]—raises a host of ethical and moral questions. For example, "Will soldiers

[25] Although there are a number of weapon systems in use today that depend on varying degrees of human supervision, there are none that are fully autonomous (with the only possible exception being the Israel Defense Forces *Harpy*, a "fire-and-forget" loitering munition

be willing to go to battle alongside robots?" "Will robots be able to distinguish between military and civilian targets, and be able to use force proportionately?" "Will an AI be able to recognize enemy signs of surrender?" "Who will be responsible for an unjustified robotic kill?" and "How does one codify an innately subjective body of ethical standards and practices?"

Such questions have led to several international movements against "killer robots."[26] For example, in July 2015, over 1,000 robotics and artificial intelligence researchers signed an open letter calling for a ban on offensive autonomous weapons (with 20K+ signatories as of Dec 2016).[27] And, at the most recent United Nations Convention on Conventional Weapons, the 123 participating nations voted to convene a group of government experts to meet (during two sessions) in 2017 to formally address the LAWS issue, which could potentially lead to an international ban.[28]

While the outcome of these upcoming meetings is uncertain, it is clear is that the political, cultural, and basic human-rights dimensions of this issue are only beginning to be explored. An analysis of the *operational* impact that any limitations on (or an outright ban of) the use of offensive autonomous weapons may entail for U.S. military forces obviously deserves attention.

Transitioning to new autonomy-enabled mission areas

Figure ES-1 illustrates, schematically, the key steps involved in extending the existing unmanned systems mission space (e.g., reconnaissance, route clearance, and search and rescue) to one that more fully embraces all that autonomy potentially offers (e.g., self-organized, and self-healing, adaptive swarms). Leaving aside details of the pipeline to the main text, the key (mutually entwined) steps include, starting from bottom of the figure and working our way to the top:

- *Step 1:* Conducting basic AI research across multiple domains (the green-to-red overlay emphasizing that research in different AI areas—e.g., deep learning,

designed to detect, attack and destroy radars). Autonomy policy for U.S. weapon systems is spelled out in DoD Directive 3000.09, which expressly prohibits use of lethal *fully* autonomous weapons, which it defines as weapon systems that, once activated, may select and engage targets without further intervention by a human. Ref: DoD Directive 3000.09, "Autonomy in Weapon Systems," Nov 2012: http://www.dtic.mil/whs/directives/corres/pdf/ 300009p.pdf.

[26] M. Wareham and S. Goose, "The Growing International Movement Against Killer Robots," *Harvard International Review*, 5 Jan 2017.

[27] http://futureoflife.org/open-letter-autonomous-weapons/.

[28] Final Document of the Fifth Review Conference, CCW, Dec 2016: http://www.reaching critical will.org/disarmament-fora/ccw/2016/revcon.

image recognition, and robotic swarms—necessarily proceeds at different rates and exists, at any one time, at different levels of maturation).

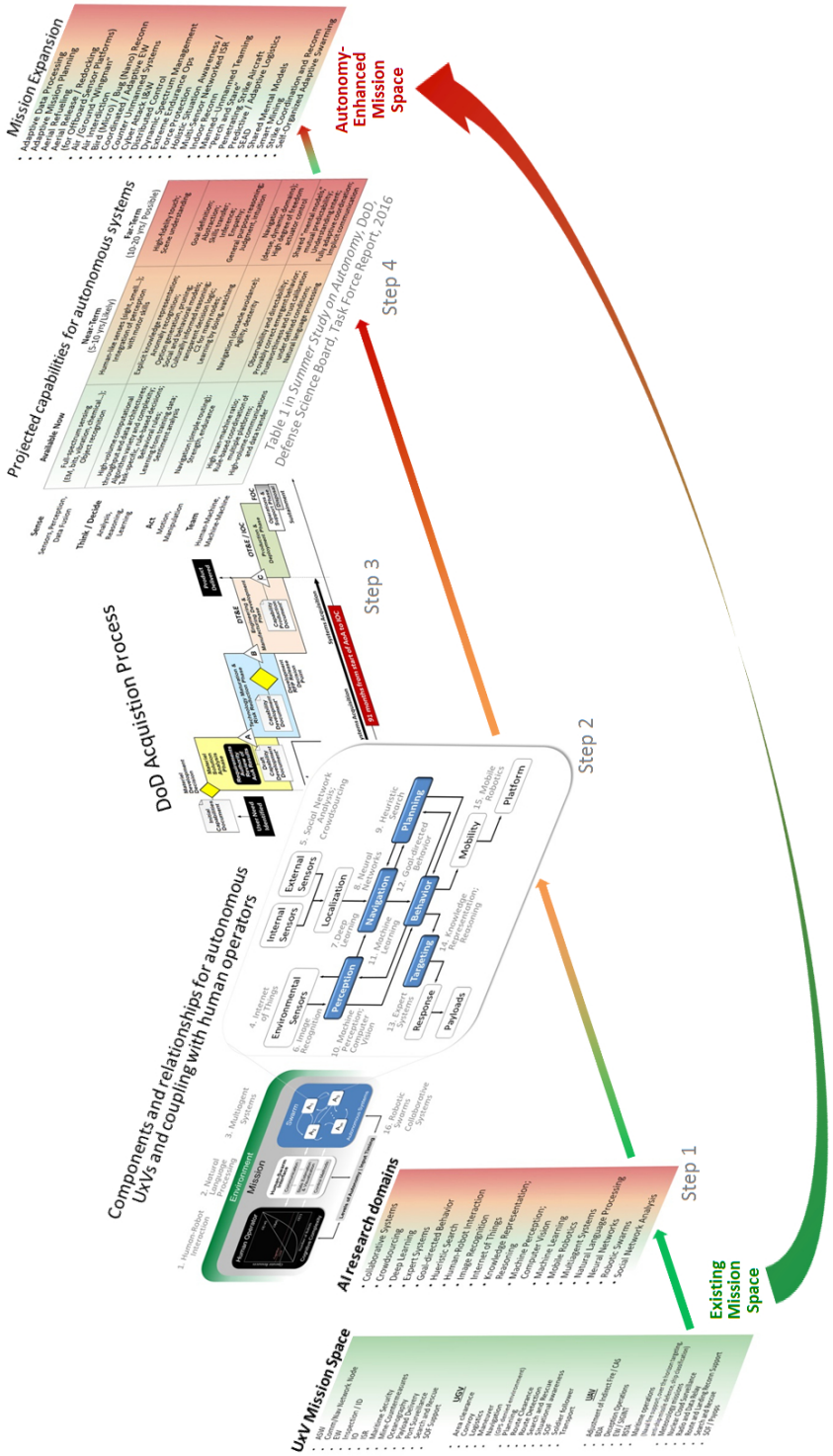
- *Step 2:* Understanding how individual AI research domains feed into the myriad components that make up autonomous systems, including their coupling with human operators (which further involves the understanding of how human-machine collaborative systems function in specific mission environments).

- *Step 3:* Moving design, development, testing, and accreditation through the DoD acquisition process (and accommodating autonomy's unique set of technical challenges while doing so).

- *Step 4:* Interpreting and projecting the requisite levels of maturity of system capabilities that autonomous systems must possess for specific missions. The autonomous systems that are shown in figure ES-1 are characterized as functions of four broad categories of AI (i.e., *sensing*, *thinking*, *acting*, and *teaming*). Their projected capabilities are indicated as follows: shades of green indicate capabilities that are available now; shades of orange denote near-term capabilities; and increasingly darker shades of red indicate the far-term regime. This table is taken from the DoD's Defense Science Board's most recent study on autonomy,[29] but is intended mostly as a notional place-holder for the kinds of conceptual, technical, and analytical considerations that must be taken into account as the raw capabilities of the autonomous systems that come out of the acquisition process are transformed into new and operationally meaningful missions and missions areas.

[29] Table 1 in *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016: https://www.hsdl.org/?view&did=79464.

Figure ES-1. Key steps in transitioning to new autonomy-enabled mission areas



Gestalt of main findings

The military is on the cusp of a major technological revolution as it enters the *Robotic Age,*[30] in which warfare is conducted by unmanned and increasingly autonomous weapon systems, operating across all domains (air, sea, undersea, land, space, and cyber), and across the full spectrum of military operations. The question is not *whether* the future of warfare will be filled with autonomous, AI-driven robots, but *when* and in what *form*. However, unlike the last "sea change" during the Cold War (i.e., the so-called "2nd Offset"),[31] when advanced technologies such as precision-strike weapons, stealth aircraft, smart weapons and sensors, and GPS were developed primarily by DoD-sponsored research and development programs, a successful transition into the Robotic Age (spurred on by DoD's recent "Third Offset Strategy" innovation initiative)[32] will depend critically on how well DoD is able to embrace technologies and innovations that are now being developed mostly in the commercial world. And, while the human warfighter is not going away anytime soon, if ever (even as the depth and breadth of autonomy steadily expand), human operators will not suddenly lose control of existing unmanned systems. A telltale sign that DoD has made a "no looking back" cross-over into the Robotic Age will be when human operators can no longer fully understand, or *predict*, how autonomous systems behave—i.e., when, for the first time, a human operator is as stunned by some weapon system's action as 18-time world Go champion Lee SeDol was by a single move of the AI that defeated him.

In preparation for DoD's cross-over into the Robotic Age, whenever it arrives, this study has identified four key technical gaps in developing AI-based autonomous systems, wherein opportunities for future analytical studies naturally arise (see figure ES-2).

These gaps are:

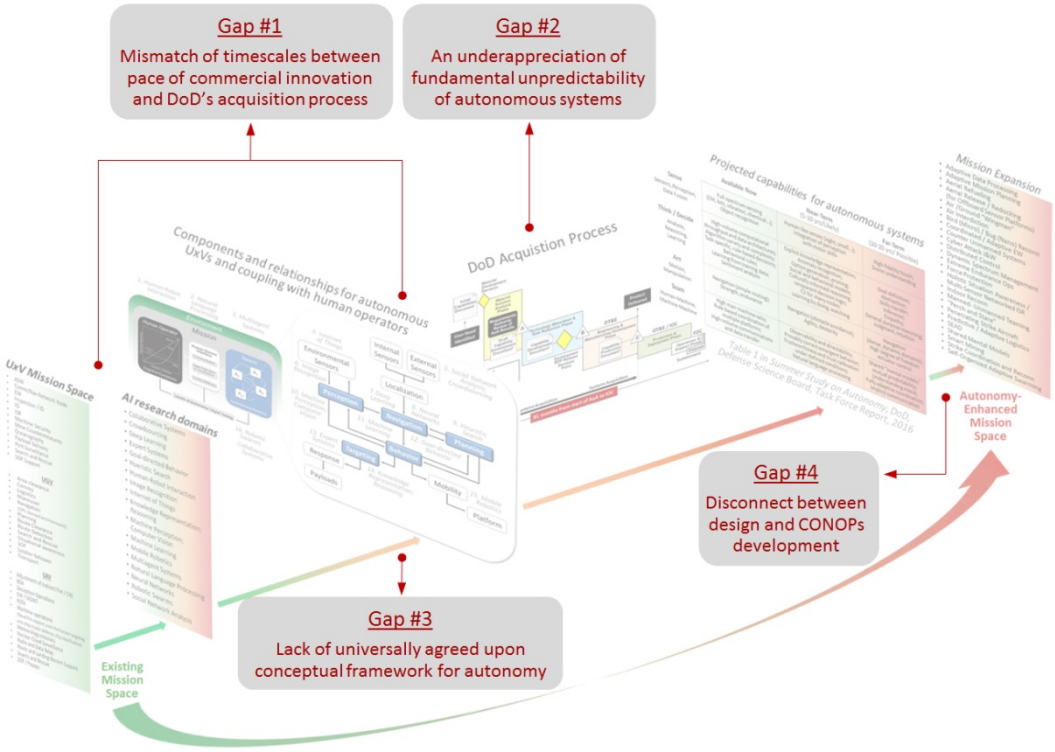
- *Gap 1*: A fundamental mismatch—even *dissonance*—between the accelerating pace (and manner of development and evolution) of technology innovation in commercial and academic research communities, and the timescales and assumptions underlying DoD's existing acquisition process.

[30] Robert O. Work and Shawn Brimley, *20YY: Preparing for War in the Robotic Age*, Center for a New American Security, Jan 2014.

[31] J. McGrath, "Twenty-First Century Information Warfare and the Third Offset Strategy," *Joint Forces Quarterly*, National Defense University, Issue 82, 3rd Quarter 2016.

[32] C. Hagel, Transcript of Keynote speech delivered at Reagan National Defense Forum Keynote, Ronald Reagan Presidential Library, Simi Valley, CA, Nov. 15, 2014.

Figure ES-2. Key *gaps* in transitioning to new autonomy-enabled mission areas



- *Gap 2*: An underappreciation of the unpredictable nature of autonomous systems, particularly when operating in dynamic environment, and in concert with other autonomous systems. Existing T&E/V&V practices accommodate neither the basic properties of autonomous systems, as expected by AI and indicated by decades of deep fundamental research into the behavior of complex adaptive systems, nor the requirements they must meet, as weapon systems (as spelled out by DoD Directive 3000.09).

- *Gap 3*: A lack of a universally agreed upon conceptual framework for autonomy that can be used both to anchor theoretical discussions and to serve as a frame-of-reference for understanding how theory, design, implementation, testing, and operations are all interrelated. A similar deficiency exists for understanding the role that trust plays in shaping a human operator's interaction with an autonomous system. The Defense Science Board's most

recent study on autonomy[33] warns that "inappropriate calibration" of trust during "design, development, or operations will lead to misapplication" of autonomous systems, but offers only a tepid definition of trust, and little guidance on how to apply it.

- *Gap 4*: DoD's current acquisition process does not allow for a timely introduction of "mission-ready" AI/autonomy, and there is a general disconnect between system design and the development of concepts of operations (CONOPS). Unmanned systems are typically integrated into operations from a *manned*-centric CONOPS point of view, which is unnecessarily self-limiting by implicitly respecting human performance constraints.

Recommended studies

While not even AI experts can predict how AI will evolve in even the near–term future (much less project its possible course over 10 or more years,[34] or predict AI's impact on the development of military autonomous systems), it is still possible to anticipate many of the key conceptual, technical, and operational challenges that DoD will face in the coming years as it increasingly turns to and more deeply embraces AI-based technologies, and fully enters the "Robotic Age." From an operational analysis standpoint, these challenges can also be used to help shape future studies:

Recommendation 1: ***Help establish dialog between commercial research and development and DoD.***

Institutions specializing in operational analysis are well suited to act as "go betweens" linking the academic and commercial research communities with military culture / operational needs. Assuming that Secretary of Defense

[33] *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016: https://www.hsdl.org/?view&did=79464.

[34] S. Armstrong, K. Sotala, and S. hÉigeartaigh, "The errors, insights and lessons of famous AI predictions – and what they mean for the future," *Journal of Experimental & Theoretical Artificial Intelligence* 26, no. 3, 2014; D. Fagella, "Artificial Intelligence Risk – What Researchers Think is Worth Worrying About," *Tech Emergence,* 20 March 2016: http://techemergence.com/artificial-intelligence-risk/. For the most recent survey of expert opinion see: V. Muller and N. Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in *Fundamental Issues of Artificial Intelligence*, edited by V. Muller, Springer-Verlag, 2016.

Ashton Carter's Defense Innovation Unit-Experimental (DIUx) program survives into the next administration,[35] operationally informed and technically knowledgeable analysts can help stakeholders better "understand" each other. Cross-fertilization with the Naval Postgraduate School (NPS) may also pay dividends.[36]

Recommendation 2: ***Develop an operationally meaningful conceptual framework for autonomy.***

For example, build on lessons learned from the National Institute of Standards and Technology's (NIST's) stalled evolution of its ALFUS (Autonomy Levels for Unmanned Systems) framework, and develop the skeleton of an idea proposed by DoD's Defense Science Board's 2012 report on autonomy.[37]

Recommendation 3: ***Develop measures of effectiveness (MOEs) and measures of performance (MoP) for autonomous systems.***

Develop a methodology by which the effectiveness of autonomous systems can be measured at all levels (e.g., developers, program managers, decision-makers, and warfighters) and across all required functions, missions, and tasks (e.g., coordination, mission tasking, training, survivability, situation awareness, and workload).

Recommendation 4: ***Use nontraditional modeling and simulation (M&S) techniques to help mitigate AI/autonomy-related dimensions of uncertainty.***

As DoD moves into the Robotic Age, M&S is moving away from "simulations as distillations" of real systems (for which M&S has traditionally been used to develop models in order to gain insights into the *real* system), to "simulation-based rules and algorithms as descriptions" of real (i.e., engineered)

[35] DIUx has been established to help facilitate the discovery and development of capabilities and technologies outside DoD's normal acquisition pipeline. Ref: https://www.diux.mil/.

[36] For example: NPS's Consortium for Robotics and Unmanned Systems Education and Research (CRUSER: https://my.nps.edu/web/cruser), and *Autonomous Systems Track* (http://my.nps.edu/web/ast).

[37] *The Role of Autonomy in DoD Systems*, DoD Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

robots and behaviors. It is here, at the cusp between exploring behaviors and prescribing rules that generate them (e.g., engineering *desired* swarm behaviors), that M&S can help mitigate some of the challenges and uncertainties of developing autonomous systems and robotic swarms. For example, while "swarm engineering" methods exist to facilitate the unique design requirements of robotic swarms, no general method exists that maps individual rules to (desired) group behavior.[38]

Multi-agent based modeling techniques[39] are particularly well suited for developing these rules, and, more generally, for studying the kinds of self-organized emergent behaviors expected to arise in coupled autonomous systems (e.g., "How sensitive is an autonomous system's behavior to changes in its physical environment?", "What new command and control architectures will be needed for robotic swarms?", and "How will the control and behavior of a swarm scale with its size and mission complexity?").

Recommendation 5: ***Apply wargaming techniques to help develop new CONOPS.***

Wargaming can be used to help identify and develop new CONOPS, apply lessons-learned from the experience of using deployed systems, explore options to counter uses of autonomy by potential adversaries, and assist in training (e.g., by exploring trust issues in human-machine collaboration). Wargames can also stimulate and nurture a more unified approach to understanding autonomous system performance and behavior, provided that they are conducted with the support and participation from across all military services and domains.

[38] I. Navarro and F. Matia, "An Introduction to Swarm Robotics," *International Scholarly Research Notes*, Vol. 2013, 2013: https://www.hindawi.com/ journals/isrn/2013/608164/.

[39] A. Ilachinski, *Artificial War: Multiagent-Based Simulation of Combat*, World Scientific, 2004. See also: A. Ilachinski, "Modelling insurgent and terrorist networks as self-organized complex adaptive systems," *International Journal of Parallel, Emergent and Distributed Systems* 27, 2012; A. Ilachinski, *AOEWSim: An Agent Based Model for Simulation Interactions Between Off-Board EW Systems and Anti-Ship Missiles*, CNA, DWP-2013-U-004757, 2013; A. Ilachinski and M. Shepko, *FAC/FIAC Simulation (FFSim): User's Guide*, CNA, Annotated Briefing, 2015.

Recommendation 6: ***Develop new T&E/V&V standards and practices appropriate for the unique challenges of accrediting autonomous systems.***

For example, help ameliorate basic gaps in testing in terms of accommodating complexity, uncertainty, and subjective decision environments, by appealing to and exploiting lessons learned from the development and accreditation practices established by the complex system theory and multiagent-based modeling research communities.

Recommendation 7: ***Explore basic human-machine collaboration and interaction issues.***

As autonomy increases, human operators will be concerned less with the manual control of a vehicle, and more with controlling swarms and directing the overall mission: "What are the operator's informational needs (and workload limitations) for controlling multiple autonomous vehicles?" "How do humans keep pace with an accelerating pace of autonomy-driven operations?" "What kinds of command-and-control relationships are best for human-machine collaboration?" "How are human and autonomous-system decision-making practices optimally integrated?" and "What data practices are key to developing shared situation awareness?"

Recommendation 8: ***Explore the challenges of force-integration of increasingly autonomous systems.***

Essentially all force-integration issues are, as yet, undetermined. They must consider not just "low hanging fruit" extensions of existing CONOPS, in which the human component is simply replaced with unmanned systems and "operational value" of human performance is scaled to accommodate "better" performance (e.g., endurance, survivability), but brainstorm heretofore nonexistent tactics, operations, and missions that fully embrace existing and anticipated future autonomous capabilities. What is the tradeoff between large numbers of simple, low-cost (i.e., "disposable") vehicles and small numbers of complex (multi-functional) ones?

The operationalization of robotic swarms, in particular, represents a heretofore largely untapped dimension of the mission space, and will require the development of new CONOPS. The swarm may be used as a radically new form of

precision coordinated "en masse" guided munition; as a self-healing area surveillance network (which includes collecting and assimilating data on an adversary's Internet-of-Things (IoT);[40] or as an adaptive distributed electronic jammer.

Recommendation 9: ***Explore the cyber implications of autonomous systems.***

Explore what new features increased AI-driven autonomy brings to the general risk assessment of increasingly autonomous unmanned systems. On one hand, autonomy may potentially reduce a force's overall vulnerability to jamming or cyber hacking. For example, communications loss over a jammed data link may be compensated for by the ability of autonomous vehicles to continue performing their mission). On the other hand, autonomy itself may also be *more*, not less, vulnerable to a cyber intrusion. For example, an adversary may gain "control," or otherwise deliberately "perturb" the behavior of an autonomous system; it may also be more difficult to detect embedded malware. In the latter context, consider some future variants of incidents such as the Iranian capture of an RQ-170 *Sentinel* in 2011,[41] and the "keylogging" virus that infected the UAV-control-computers at the Creech Air Force Base in Nevada.[42]

Recommendation 10: ***Explore operational implications of ethical concerns over the use of lethal autonomous weapon.***

Analyze issues of accountability, legality, and liability in arguments put forth by various "Ban LAWS" movements. Examine the possible constraints on missions (along with other associated impediments to the design and development of autonomous systems) that may result from an international ban (or set of limits) imposed on the development or deployment of LAWS, such as might come out of the United-Nations-sponsored government experts' negotiations scheduled to take place sometime in 2017.

[40] G. Seffers, "Defense Department Awakens to Internet of Things," *Signal*, 1 Jan 2015: http://www.afcea.org/content/?q=defense-department-awakens-internet-things.

[41] The Iranian government announced that the RQ-170 was captured by its cyber warfare unit: "Iran shows film of captured US drone," BBC News, 8 Dec 2011: http://www.bbc.com/news/world-middle-east-16098562.

[42] N. Shachtman, "Exclusive: Computer virus hits U.S. drone fleet," *Wired*, 7 Oct 2011.

Contents

This page intentionally left blank.

List of Figures

List of Tables

This page intentionally left blank.

Glossary

AFRL	Air Force Research Laboratory
AoA	Analysis of Alternatives
ACO	Ant Colony Optimization
A2/AD	Anti-access and area denial
ABC	Artificial Bee Colony
AI	Artificial Intelligence
AoA	Analysis of Alternative
ASRF	Autonomous System Reference Framework
AUV	Autonomous Undersea Vehicle
AWS	Autonomous Weapon System
ALFUS	Autonomy Levels for Unmanned Systems
BL	Backpropagation Learning
CSKR	Campaign to Stop Killer Robots
CNAS	Center for a New American Security
CEC	Cognitive Echelon Class
CNN	Convolutional Neural Network
C2	Command and Control
CONOPs	Concept of Operations
CAC	Contextual Autonomous Capability
CARACaS	Control Architecture for Robotic Agent Command Sensing
CAS	Complex Adaptive System
CPS	Computations Per Second
DAMS	Defense Acquisition Management System
DARPA	Defense Advanced Research Projects Agency
DL	Deep Learning
DSB	Defense Science Board

DIUx	Defense Innovation Unit-Experimental
DoD	Department of Defense
DOTMLPF	Doctrine, Organization, Training, Materiel, Leadership and Education, Personnel, and Facilities
EMD	Engineering & Manufacturing Development
DAP	DoD Acquisition Process
DoDD	DoD Directive
EC	Environmental Complexity
GA	Genetic Algorithm
GAO	Government Accountability Office
HUMS	Health and Usage Monitoring System
HI	Human Interface
HMSTSC	Human-Machine System Trades Space Class
HRI	Human-Robot Interaction
IED	Improvised Explosive Device
IT	Information Technology
IOC	Initial operational Capability
IEEE	Institute of Electrical and Electronics Engineers
INTEL	Intelligence
ISR	Intelligence, Surveillance, and Reconnaissance
ICRAC	International Committee for Robot Arms Control
IHL	International Humanitarian Law
LSVRC	Large Scale Visual Recognition Challenge
LAWS	Lethal Autonomous Weapon System
LPF	Linear Predictor Function
LR	Linear Regression
MUM	Manned-Unmanned
MSA	Materiel Solution Analysis
MC	Mission Complexity
MDC	Mission Dynamics Class
MBM	Multiagent-Based Model
NAT	Normal Accident Theory
NDAA	National Defense Authorization Act

NSS	National Security System
NLP	Natural Language Processing
NLU	Natural Language Understanding
NN	Neural Network
NECSI	New England Complex Systems Institute
NIST	National Institute of Standards and Technology
NGO	Non Governmental Organization
OODA	Observe, Orient, Decide, and Act
ONR	Office of Naval Research
O&S	Operations and Support
PD	Production and Deployment
RS	Robotic Swarm
SFI	Santa Fe Institute
S&T	Science and Technology
SOC	Self Organized Criticality
SCO	Strategic Capabilities Office
SME	Subject Matter Expert
SI	Swarm Intelligence
SoS	Systems of Systems
TF	Task Force
TMRR	Technology Maturation & Risk Reduction
T&E	Test and Evaluation
TOS	Third Offset Strategy
TRoR	Three Rules of Robotics
TL	Turing Leaning
UAV	Unmanned Aerial Vehicle
UAS	Unmanned Autonomous System
UGV	Unmanned Ground Vehicle
USIR	Unmanned System Integrated Roadmap
UUV	Unmanned Underwater Vehicle
V&V	Verification and Validation

This page intentionally left blank.

Introduction

The industrial commercial sector is currently undergoing a dramatic transformation due to the rapid growth of robotics and other artificial intelligence (AI) technologies, from drones, to self-driving cars, to virtual reality, to wearable devices, to human brain-to-brain interface technologies, to deep-learning machine learning techniques.[1]

Twenty years ago, in 1997, *Twitter* and *Facebook* did not yet exist,[2] and *Google* had just appeared on the technological landscape.[3] Back then, the human genome had not yet been sequenced; today, commercial genome sequencing products are available for individual purchase.[4] In AI, IBM's *DeepBlue* chess computer had just defeated Gary Kasparov, the (then) reigning chess champion;[5] in 2016, Google's *AlphaGo* program beat one of the highest ranked players in the world at the game of Go (a game that is so much more "complex" than chess as a problem for an AI to "solve," that prior to *AlphaGo*'s victory, most AI experts believed that it could not be done for another 15-20 years).[6] Self-driving cars were confined to basic research programs in 1997,[7] but Google's prototype appeared in 2009[8] and, by the end of 2016, the total number of miles driven by Tesla's first-generation commercially available *Autopilot*-

[1] J. Kadtke and L. Wells II, *Policy Challenges of Accelerating Technological Change: Security Policy and Strategy Implications of Parallel Scientific Revolutions*, Center for Technology and National Security Policy, National Defense University, Sep 2014.

[2] S. Edosomwan, "The History of Social Media and its Impact on Business," *The Journal of Applied Management and Entrepreneurship* 16, no. 3, 2011.

[3] https://www.google.com/about/company/history/.

[4] *Science Exchange* lists 37 labs offering services (as of this writing): https://www.scienceexchange.com/services/whole-genome-seq.

[5] M. Campbell, A. Hoane, Jr., and F. Hsu, "Deep Blue,"*Artificial Intelligence* 143, 2002.

[6] C. Koch, "How the Computer Beat the Go Master," *Scientific American*, 19 March 2016.

[7] Carnegie Mellon University's *Navlab* project semi-autonomously completed a 3,100 mile cross-country in 1995: http://www.cs.cmu.edu/afs/cs/usr/tjochem/www/nhaa/nhaa_home_page.html.[8] M. Harris, "How Google's Autonomous Car Passed the First U.S. State Self-Driving Test,"*IEEE Spectrum*, 10 Sep 2014.

[8] M. Harris, "How Google's Autonomous Car Passed the First U.S. State Self-Driving Test,"*IEEE Spectrum*, 10 Sep 2014.

driven cars had exceeded 300 million miles.[9] And, while simple radio-controlled aircraft were in the hands of amateur hobbyists back in 1997,[10] it was only in the last decade or so that today's cheap, lightweight, quadcopters (which use advanced electronics for stability and flight control) entered the market.[11] The market volume for consumer drones was over 6 million units in 2015, and is projected to increase *tenfold* to 67.9 million by 2021.[12] The total global market for commercial applications of UAS technology is estimated to go over $125 billion by 2020 (compared with $2 billion today).[13]

A recent study of the future of robotics and AI concluded that the emergence of "intelligent machines" will essentially define the next industrial revolution, estimating that smart machines and robots will perform 45% of all manufacturing tasks by 2025 (compared to about 10% in 2016).[14] Global spending on commercial and industrial robotics is estimated to be over $43 billion per year by 2018.[15] A recent study conducted by Bank of America/Merrill Lynch projects that the total AI market (valued at $2 billion in 2015), will grow to $36 billion by 2020, and to $127 billion by 2025.[16]

The military is certainly a part of these recent developments. For example, even relatively "old" technologies such as the *Predator* (which entered production in

[9] F. Lambert, "Tesla has now 1.3 billion miles of Autopilot data going into its new self-driving program," *Electrek,* 13 Nov 2016.

[10] The first company to offer model airplane kits for sale was Radioplane Co. Inc., founded in 1935 by Reginald Denny B. Benchoss, "A Brief history of the drone," *Hackaday*, 26 Sep 2016.

[11] The personal drone movement can be traced to: (1) an online community and forum called *DIYDrones.com*, and founded in 2007 by Chris Anderson (future CEO of 3D Robotics), where technically minded hobbyists could meet, share ideas, and discuss building drones; and (2) the *DJI Phantom*, the first small quadcopter that was desigend specifically for consumers and put on sale in 2014 (C. Guillot, "Commercial drones," *Business Researcher*, 18 Jan 2016).

[12] "Consumer Drone Sales to Increase Tenfold to 67.7 Million Units Annually by 2021," *Tractica*, 6 July 2016.

[13] *Rise of the Drones*, Allianz, 2016: www.agcs.allianz.com/assets/PDFs/Reports/AGCS_Rise_of_the_drones_report.pdf.

[14] "Robot Revolution: Global Robot an AI Primer," *Thematic Investing*, Bank of America and Merrill Lynch, 16 Dec 2015: https://www.bofaml.com/content/dam/boamlimages/documents/PDFs/robotics_and_ai_condensed_primer.pdf.

[15] M. Horowitz, "The Looming Robotics Gap," *Foreign Policy,* May 5, 2014: http://www.foreignpolicy.com/articles/2014/05/05/the_looming_robotics_gap_us_military_technology_dominance.

[16] *Thematic Investing: Robot Revolution—Global Robot & AI Primer*, Bank of America/Merrill Lynch, 16 Dec 2015.

1997)[17] and *Packbot*[18] ground robot (first deployed in 2002) have already demonstrated their value. However, what these technologies represent to the military as a whole (and the Navy, in particular)—tactically, operationally, and as part of a roadmap that can potentially redefine how warfare is conducted in both near-term and far-term futures (toward "war in the robotic age"[19])—has not yet been seriously addressed on an analytical level. Arguably, autonomous weapon systems—and related hybrid robot-human technologies, interfaces, and other combined biological and electronic systems—are already transforming a "merely" networked force into a genuine swarm, and, collectively, have the potential to be as disruptive to conventional military wisdom and practice, as digital data links once were a generation or so ago. As burgeoning anti-access weapons curtail traditional channels of power projection (and non-state actors themselves deploy "robot swarm" weaponry),[20] autonomous systems may, sooner rather than later, be the only "go to" technology that reliably penetrates the enemy space. However, the global spending on *military* robotics is estimated to reach only $7.5 billion per year by 2018,[21] so that—when compared to the $43 billion per year estimate quoted above for the commercial sector—it can be anticipated that much of the near-term progress will not, as in priori generations, come from the U.S. defense sector.

The Navy is laudably developing new "swarming" technologies. Two recent examples are the Office of Naval Research's LOCUST = Low Cost (Unmanned Aerial Vehicle (UAV) Swarm Technology,[22] and a 2014 demo of CARACaS = Control Architecture for Robotic Agent Command Sensing, in which 13 unmanned surface vessels escorted a manned control ship through the James River.[23] (Many more appear throughout this white paper.) However, the concomitant operational analysis needed to understand the potentially revolutionary impact of these technologies on future military operations is lagging.

[17] http://www.airforce-technology.com/projects/predator-uav/.

[18] http://www.army-technology.com/projects/irobot-510-packbot-multi-mission-robot/.

[19] S. Brimley and P. Scharre, "Time to Get Ready for War in the Robotic Age," *Defense One*, 26 Jan, 2014.

[20] R. Bunker, Terrorist and Insurgent Unmanned Aerial Vehicles: Use, Potentials, and Military Implications, U.S. Army War College, Strategic Studies Institute, August 2015.

[21] P. Scharre, *Robotics on the Battlefield Part II: The Coming Swarm*, Center for a New American Security, Oct 2014.

[22] http://www.onr.navy.mil/Media-Center/Press-Releases/2015/LOCUST-low-cost-UAV-swarm-ONR.aspx.

[23] http://www.onr.navy.mil/Media-Center/Press-Releases/2014/autonomous-swarm-boat-unmanned-caracas.aspx.

Organization of this report

This report summarizes the results of an "exploratory" study of the state-of-the-art of AI, machine-learning, and robot technologies, their potential future military implications for autonomous (and semi-autonomous) weapon systems, and the development of general autonomy-centric concept of operations (CONOPs) and swarm-vs-swarm combat tactics.

The study has three main goals:

1. Identify (and provide a context for) key technical issues and analysis gaps.
2. Provide a roadmap of opportunities and challenges.
3. Recommend potential future studies.

The main body of the report may be viewed as a prolonged evidence-based written "argument" justifying its core thesis (defined below), and culminating with a list of study recommendations. It is structured, thematically and organizationally, around a set of basic questions:

- What is the state-of-the-art in artificial intelligence and robotics technology?
- What is a swarm? Can it be reliably and predictably "controlled"? How?
- What potential impacts will robotic swarm technology have on military operations?
- Will robot-swarm technology gracefully enfold within (and around) traditional mission areas, or will it fundamentally transform the nature of warfare?
- What does "autonomy" really mean?
- What are the tradeoffs between human-based and autonomous systems?
- What are the tradeoffs between the operational benefits and risks of autonomous systems?
- Can DoD's current acquisition process accommodate the accelerating pace of innovations and developments in AI?
- Are traditional forms of Test and Evaluation (T&E) and Verification and Validation (V&V) adequate for autonomous systems?
- Will existing command-and-control (C2) architectures have to be changed in order to accommodate various levels of autonomy?
- What are the ethical and moral questions and risks associated with autonomous weapon systems? (The most egregious is the spectre of the

"Frankenstein" scenario—i.e., the "…haunting fear that [we] will be unable to control what [we] create."[24])

- What new forms of military operations research will be required to address autonomous weapons systems?

Each of these questions is addressed—and answered, to varying degrees—in the sections that follow. Of course, given the vast scope of the subject matter of this report, it is impossible, even in a document as lengthy as this one, to cover all pertinent issues with equal depth. It is our hope that, at the very least, the document will serve as a stepping stone for analysts interested in AI, robots, and autonomy to pursue their own research into these subjects.

Apart from providing specific recommendations on how military operations research analysis, in general, and the Center for Naval Analyses (CNA), in particular, can help the Navy (and DoD) navigate—and better understand the conceptual and technical dimensions of—a path toward increasing levels of autonomy, this report is also intended to serve as a general "go to" sourcebook of information about AI and swarm technologies as they relate to military operations, and to provide a summary of the recent spate of DoD directives, memos, and task force reports on the future role of autonomy. The discussion, arguments, and assertions that appear herein are supported by a bibliography containing over 150 references to original source papers and research material, and over 725 footnotes (that contain several hundred links to additional online sources of information and data). The appendix contains a set of slides that summarize a selection of recent "pushing the envelope" technological innovations (which have been made both inside and outside of the military).

Terminology

Much confusion can (and does) arise in the literature when terms such as "autonomy," "automated," and "robot" are used inconsistently, at best, and incorrectly, at worst. To avoid unnecessary confusion, we begin our discussion by introducing basic nomenclature (which will be revisited throughout the report) :[25]

[24] T. Simpson, "Robots, trust, and war," *Philosophy and Technology* 24, May 2011.

[25] A. Finn and S. Scheding, *Developments and Challenges for Autonomous Unmanned Vehicles: A Compendium*, Springer-Verlag, 2010; R. Mittu, D. Sofge, A. Wagner, and W. Lawless, editors, *Robust Intelligence and Trust in Autonomous Systems*, Springer-Verlag, 2016; R. Murphy, *Introduction to AI Robotics*, MIT Press, 2000.

- *Artificial Intelligence* (AI): The ability of a system "devoted to making machines intelligent," where intelligence is that "quality that enables an entity to function appropriately and with foresight in its environment."[26]
- *Intelligent System* (IS): An application of AI to a particular problem domain (referred to as "narrow AI"). It is typically very specialized, though increasingly capable of "super human" capability (see discussion in a later section), and does not represent "general intelligence" (or "general AI").
- *Automated System*: A physical system that functions with no (or limited) human operator involvement, typically in structured and unchanging environments, and whose performance is limited to the specific set of actions it has been designed to accomplish. Typically these are well-defined tasks that have predetermined responses (i.e., behaviors are "scripted" according to simple rule-based prescriptions).
- *Autonomous System* (AS): An IS that is able to independently compose and select among alternative courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the local, dynamic context. Unlike automated systems, autonomous systems are designed to respond to situations that are not pre-programmed or anticipated prior to system deployment.

  There are three broad classes of autonomy as it pertains specifically to the role it plays in the use weapons, delineated by the degree of control that humans can exert on the weapon deployment:[27]

  - *Semi-autonomous* ("human in the loop"): Once the weapon system is activated, it engages only those targets that have been selected by a human operator. (Specific functions may include acquiring, tracking, and identifying, cueing, and prioritizing potential targets.)
  - *Supervised autonomous* ("human on the loop"): Once activated, the weapon system operates under human supervision. (That is, the human operator can intervene and terminate engagements, including in the event of a weapon system failure.)
  - *Fully autonomous* ("human out of the loop"): Once activated, the weapon system may select and engage targets without further

[26] N. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge University Press, 2009. Note that there is no universally agreed-upon definition of what "AI" is. The definition provided here is intended only to "set the stage" for discussion.

[27] DoD Directive 3000.09, *Autonomy in Weapon Systems*, Nov 2012: http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

intervention by a human. (No human either supervises the operation of the task nor has an ability to intervene in the event of a system failure.)

- *Robots*: Physical systems with on-board sensors and actuators that are able to operate autonomously or semi-autonomously in cooperation with humans. Basic robotics research (examples of which are given later) focuses on developing adaptive intelligence to enable robots to cope with unstructured dynamic environments.[28]
- *Agent*: A self-activating, self-sufficient and persistent computation. It may be an IS; may include significant automation; may be able to modify the manner in which it achieves the objective; and may reside and act entirely in the cyber world, or be embodied in a physical system such as a robot.

Of course, the intuitive "simplicity" of most of these definitions deceptively belies the multiple nested "devil is in the details" layers of complexity that must be dealt with, and from which, as will be argued, come both opportunities and challenges.

Core thesis

This study is predicated on three tenets (the justification for which, and meaning of, will be provided and amplified during the ensuing discussion):

- *Tenet-1*: That the exponential growth of *affordable* computing power, measured in terms of raw calculations per second, may soon—possibly within 10-15 years—achieve a relative parity with that of the human brain (~ 10 petaflops = 10^{16} computations per second (CPS) = 10,000 trillion CPS);[29]
- *Tenet-2*: That *limited domain* AI—that is, AI applied to "solving" relatively narrowly focused but fundamental "problems" such as image and speech recognition, trivia, chess, and Go (tasks that until fairly recently were

[28] The term "robot" is based on the Czech word *robota* (meaning "serf or slave") and was introduced as a broad cultural lexicon in Karel Capek's 1921 play *R.U.R.* (Rossum's Universal Robots). In the beginning of the play, robots are synthetic humans that work in factories to produce low-cost goods. The play ends as these robots kill off the human race.

[29] China has reportedly far surpassed the 10 petaflop level earlier this year, having achieved 93 petaflops in its *Sunway TaihuLight* supercomputer: M Feldman, "China Tops Supercomputer Rankings with New 93-Petaflop Machine," *Top 500*, 20 June 2016: https://www.top500.org/news/china-tops-supercomputer-rankings-with-new-93-petaflop-machine/.

regarded as "too hard" for current-generation AI)—already regularly outperforms humans.[30]

- *Tenet-3*: Which rests on the observation that, although DoD's total drone-related spending remains high (e.g., $5.8 billion in FY16 and $4.6 billion in FY17),[31] the procurement of new systems (specifically, of medium-sized and larger UAVs: MQ-1, RQ-4, MQ-8, MQ-9, RQ-11) is *declining*: from 1,211 in FY12, to 288 in FY13, 54 in FY14, …, and 31 in FY17).[32] Specifically, the third tenet is that there are two drivers for this reduction: (1) that purchases prior to 2012 were in direct response to high demand for Iraq and Afghanistan (and that declining requirements are met by the current generation), and (2) that, generally speaking, the technology behind the drones that were being procured in large numbers prior to 2012 was developed mainly by industry, which the military started buying in droves after recognizing its utility.[33]

Assuming that all three tenets are valid, this study's core thesis is that these tenets, collectively, provide an enormous opportunity for military operations research analysis (MORA):

- To help the DoD better understand emerging new AI– and robot–related technologies (e.g., by developing taxonomies and conceptual frameworks within which otherwise overly complex "systems" can be systematically examined);
- To provide an impartial bridge between military requirements and the products of commercial and academic research (e.g., MORA organizations can be used as "go betweens" linking DoD and private technology industries);[34] and

[30] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2016.

[31] DoD FY17 President's Budget Proposal, Release No: NR-046-16, Feb. 9, 2016.

[32] Ibid.

[33] For example, the Gnat-750 drone (a few prototypes of which were built by General Atomics in 1980s), was bought by the CIA during the Clinton administration because it was looking for a surveillance platform to use in Bosnia. While the Gnat-750 was plagued by bugs and did not perform well, its "value" was duly noted, and subsequently led to the development of its direct follow-on, the *Predator*. Ref: R. White, "The Man Who Invented the Predator," *Wired*, April, 2013.

[34] A recent example of a DoD-sponsored laison effort is the Defense Innovation Unit-Experimental (DIUx) office, established in 2015 in Sunnyvale, California, by Secretary of Defense Ashton Carter. DIUx's charter is to seek out innovative technologies and talent from Silicon Valley. Ref: M. Eaglen, "Tech-challenged Pentagon searches for a Silicon ally," American Enterprise Institute, 1 Feb 2016: https://www.aei.org/publication/tech-challenged-pentagon-searches-for-a-silicon-ally/.

- To help develop new methods and analysis techniques, and new measures of effectiveness and performance, for studying unmanned systems with increasing levels of autonomy (beyond the "low hanging fruit" variety of assessing basic tradeoffs between manned and unmanned systems, calculating optimal paths, or developing targeting algorithms).

History of AI/robotics/swarm technologies and unmanned weapon systems

The entwined history of AI, robotics, and swarm-related technologies (both software and hardware variants) is both too long and too rich for us to provide anything but a brief summary of selected milestone events. Here is a list of the major "takeaways":

- The military roots of unmanned systems go back over 100 years.
- There is generally a long period (10 to 15+ years) of gestation before a technology matures and is available for widespread use.
- The seeds of autonomy were already in place in the early 1990s.
- Cold war era military innovations (stealth, precision navigation, satellites, networking, etc.) were generally spurred by government research and development.
- There is an accelerating pace of technology innovation.
- There is a widening (though opportunistic) "analysis gap" between requirements and technology.
- There are increasingly complex tradeoffs between controllable and unanticipated behaviors.
- The technical enablers of 20XX-era (e.g., autonomy, robotics, big data, deep learning, etc.) are driven primarily by the commercial world and the academic research community.

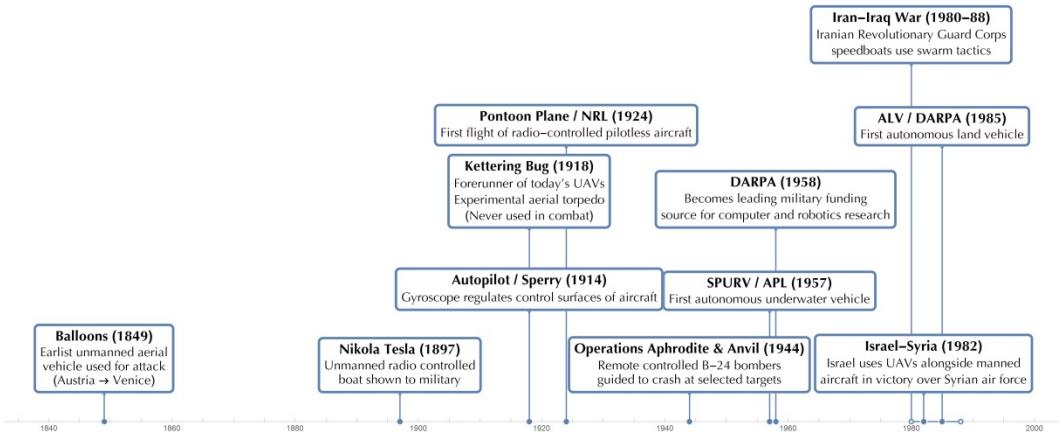
These takeaways will be amplified by the commentary that appears in this section, and more deeply supported by discussions of specific topics as they arise throughout the main narrative.

Timeline of unmanned systems

Unmanned military weapon systems have a long (and often surprisingly rich) history that dates back over 150 years, to Austria's use of pilotless balloons to drop bombs on Venice in 1849,[35] and encompasses both world wars.[36]

Figures 1 and 2 show timelines of selected milestone events in the development and deployment of unmanned systems. Figure 1 covers the "early" period, through the 1980s Iran-Iraq war, and figure 2 covers more recent events.

Figure 1. Timeline of selected milestones in the development and use of military unmanned systems (from 1849 to 1988)



The first military use of a UAV dates back to World War I, when, in 1917, the UK attempted (unsuccessfully) to use a radio-controlled Sopwith Camel biplane loaded with dynamite to dive-bomb a German zeppelin.[37] In World War II, attempts (also, largely unsuccessful) were made to develop unmanned B-17 and B-24 bombers to dive into German military-industrial targets. U.S. fighter pilots used some of the earliest drones for target practice (e.g., Dennymite, a radio-controlled plane invented

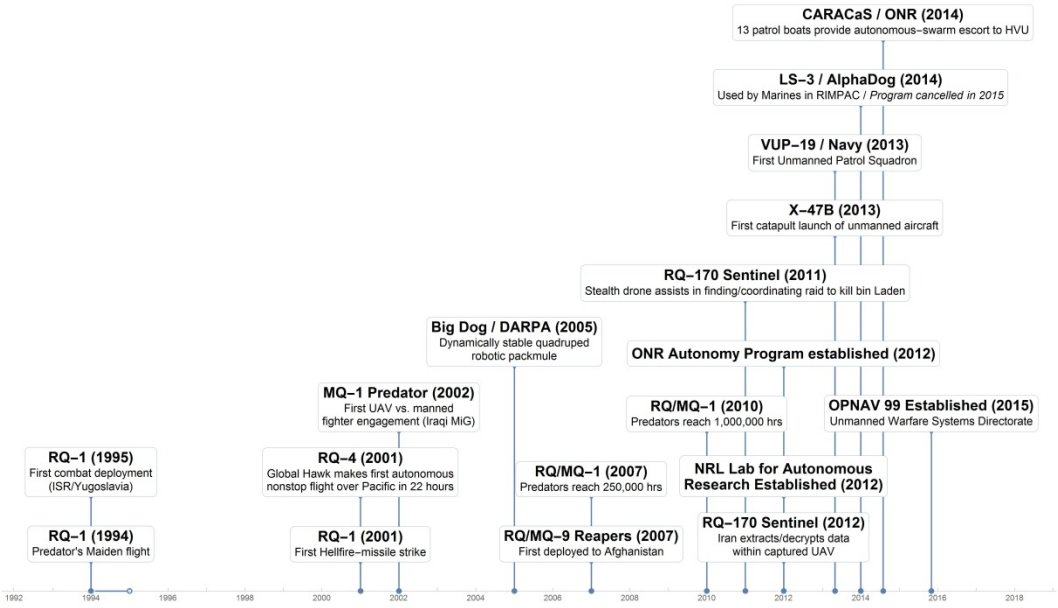
[35] L. Hargrave, *Remote Piloted Aerial Vehicles: An Anthology*, Australia's National Library in Canberra: http://www.ctie.monash.edu/hargrave/rpav_home.html#Beginnings.

[36] H. Everett, *Unmanned Systems of World Wars I and II*, MIT Press, 2015.

[37] P. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, Penguin Books, 2009.

by a British World War I pilot, Reginald Denny).[38] And Germany's V-1 "is arguably one of the first widely used "autonomous" weapons (over 8,000 so-called "flying bombs" were used).[39]

Figure 2. Timeline of selected milestones in the development and use of military unmanned systems (1988 to present day)



During the Cold War, the U.S. military's interest in unmanned system was marginal, and focused on the reconnaissance mission (e.g., the AQM-91 "Fire Fly" drones were developed during Vietnam War to fly over China).[40] A low point was the U.S. Army's Aquila program, which was launched in 1979 with the goal of developing a remotely operated drone that could be used to provide surveillance over enemy territory. But

[38] Ibid., p. 49. Denny also founded Radioplane Co. Inc., the first company to offer model airplane kits for sale: B. Benchoss, "A Brief history of the drone," *Hackaday*, 26 Sep 2016. The word "drone" was first used to describe aerial vehicles when the sound of low-flying biplanes in the 1930s was likened to the sound of a cloud of bees (and, in Old English, "drone" denotes a male honeybee: B. Zimmer, "The Flight of 'Drone' From Bees to Planes," *The Wall Street Journal*, 26 July 2016.

[39] P. Springer, *Military Robots and Drones: A Reference Handbook*, ABC-CLIO, 2013.

[40] "Teledyne Ryan AQM-91 Firefly," *Directory of U.S. Military Rockets and Missiles:* http://www.designation-systems.net/dusrm/m-91.html.

the program was canceled eight years—and about $1 billion dollars—later with only a few prototypes ever built.[41]

The turning point with regard to the military's interest in UAVs arguably is marked by the Israeli Air Force's coordinated use of manned and unmanned aircraft in its victory over the Syrian Air Force in 1982 (in which 86 Syrian aircraft were destroyed over the Bekaa Valley with minimal Israeli losses).[42] Israeli drones provided mainly real-time surveillance, but were also used as electronic decoys and jammers. These early Israeli drone successes arguably jump-started the U.S. military acquisition of UAVs (e.g., the Hunter RQ-5A / MQ-5B/C UAVs derive from Israeli models).

By the First Gulf War (which may be considered the first "UAV war," in which over 500 sorties and 1,640 hours were logged with UAVs),[43] "at least one UAV was airborne at all times during Desert Storm."[44] By way of comparison, the total number of UAV hours flown in the Second Gulf War is estimated to be over 500,000.[45] Also, notably, since the First Gulf War there has not been a conflict where UAVs were not deployed.

Other milestones include:

- Confederate and Union forces both flew balloons for reconnaissance missions during the civil war. (Though these balloons were obviously not "unmanned systems," they provide a benchmark for the number of years it took to progress from manned to unmanned flights for surveillance.)
- Aerial surveillance emerged during the 1898 Spanish–American War, when the U.S. military deployed a camera on a kite, thereby producing the world's first aerial reconnaissance photos.[46]
- The first demonstrations of a remote control by radio were orchestrated in the late 1890s by Nikola Tesla, culminating in an exhibition in 1897 at

[41] P. Singer, *Wired for War*, p. 55.

[42] M. Dobbing and C. Cole, Israel and the Drone Wars: Examining Israel's Production, Use and Proliferation of UAVs, Drone Wars UK, Jan 2014.

[43] J. Coyne, *Airpower in the Gulf*, Aerospace Education Foundation, 1992.

[44] Ibid.

[45] L. Baldor, "Military use of unmanned aircraft soars," *Associated Press*, January 2 2008.

[46] N. Polmar et al., *Spyplanes: The Illustrated Guide to Manned Reconnaissance and Surveillance Aircraft from World War I to Today*, Voyageur Press, 2016.

Madison Square Garden, New York, when Tesla caused a small boat to obey commands from the audience.[47]

- The first functioning unmanned aerial vehicle (UAV)—the *Kettering Bug*—was developed in 1918 by Orville Wright and Charles F. Kettering.[48] It was a bit over 12 feet long, four-and-a-half feet high, weighed 530 pounds (loaded), had a maximum speed of 120 mph, and a range of 75 miles. Its armament was 180 pounds of high explosive.

- Operations Aphrodite and Anvil were code names of the U.S. Air Force and U.S. Navy operations, respectively, to use B-17 and PB4Y bombers as precision-guided munitions against bunkers and other hardened/reinforced enemy facilities during World War II.[49]

- Iranian Revolutionary Guard Corps speedboats use swarm tactics for first time during the Iran-Iraq war (1980-1988).[50]

- The Predator's maiden flight took place in 1994, with its first combat deployment (as an intelligence, surveillance and reconnaissance (ISR) platform) in Yugoslavia the following year.[51]

- The first drone strike by the United States was a Hellfire-missile attack by a CIA Predator on October 7, 2001, in an unsuccessful attempt to kill Taliban Supreme Commander Mullah Mohammed Omar.[52]

- The first air-to-air engagement between a UAV and a manned aircraft took place in December 2002, between an Iraqi MiG-25 and an American Predator UCAV armed with Stinger missile.[53] (Both vehicles fired at each other, and the UAV was shot down.)

---

[47] A. Marincic, "Tesla's multi-frequency wireless radio controlled vessel," in *History of Telecommunications Conference*, 2008: http://ieeexplore.ieee.org/document/4668708/

[48] Kettering Aerial Torpedo "Bug," National Museum of the Air Force, 7 April 2015: http://www.nationalmuseum.af.mil/Visit/MuseumExhibits/FactSheets/Display/tabid/509/Article/198095/kettering-aerial-torpedo-bug.aspx.

[49] J. Olsen, *Aphrodite: Desperate Mission*, I Books, 2014.

[50] F. Haghshenass, *Iran's Asymmetric Naval Warfare,* The Washington Institute for Near East Policy, Policy Focus #87, September 2008.

[51] R. Whittle, *Predator: The Secret Origins of the Drone Revolution*, Henry Holt and Co., 2014.

[52] C. Woods, "The Story of America's Very First Drone Strike," *The Atlantic*, 30 May, 2015.

[53] D. Fulghum, "Predator's Progress," Aviation Week & Space Technology, March 3, 2003; M. Knights, *Cradle of conflict: Iraq and the birth of modern U.S. military power*, Naval Institute Press, 2005.

- *BigDog* is a dynamically stable quadruped robot developed by Defense Advanced Research Projects Agency (DARPA) in 2005, and represents a technological breakthrough in legged robotics.[54] It is essentially a robotic packhorse, with an on-board computer that controls locomotion, processes sensors (including those for joint position, joint force, ground contact, ground load, a gyroscope, and a stereo vision system), and handles communications with the user. *BigDog* runs at 4 mph, climbs slopes up to 35 degrees, walks across rubble, climbs muddy hiking trails, walks in snow and water, and carries a 340-pound load.[55] *BigDog*'s descendent, the LS3 (prototypes of which were demonstrated in 2012,[56] and used in RIMPAC 2014[57]), can carry up to 400 pounds of gear and enough fuel for a 20-mile mission lasting 24 hours, and can automatically follow its leader via computer vision. The LS3 program was canceled in 2015 due to several limitations: its loud noise, its difficulty traversing certain terrains, and the general challenge of integrating it into a traditional Marine patrol (it was used mainly as a logistical device, rather than, as originally planned, a tactical one).[58]

Proliferation of drones

Not surprisingly, the September 11, 2001, terrorist attacks initiated a general flurry of advances in military technology related to unmanned systems, and caused the use of drones to grow dramatically. For example, the Air Force logged its first 250,000 hours of drone flight time between 1995 and May 2007. But the next 250,000 hours of drone flight time took only a year and a half (from May 2007 to November 2008); and a third batch of 250,000 flight-time hours took just one year (December 2008 to December 2009).[59] A July 2012 report by the U.S. Government Accountability Office

[54] Before *BigDog*, most legged robots were either humanoid in design or patterned after insects. With its four legs, *BigDog* offers greater stability than a humanoid, and is able to carry heavier loads. M. Raibert, et al., "BigDog, the Rough-Terrain Quadruped Robot," *IFAC Proceedings Volumes*, vol. 41, no. 2, 2008: http://www.sciencedirect.com/science/article/pii/S1474667016407020.

[55] http://www.bostondynamics.com/robot_bigdog.html.

[56] http://www.bostondynamics.com/robot_ls3.html.

[57] S. Dietz, "Meeting LS3: Marines experiment with military robotics," *Marine Corps News*, 16 July 2014.

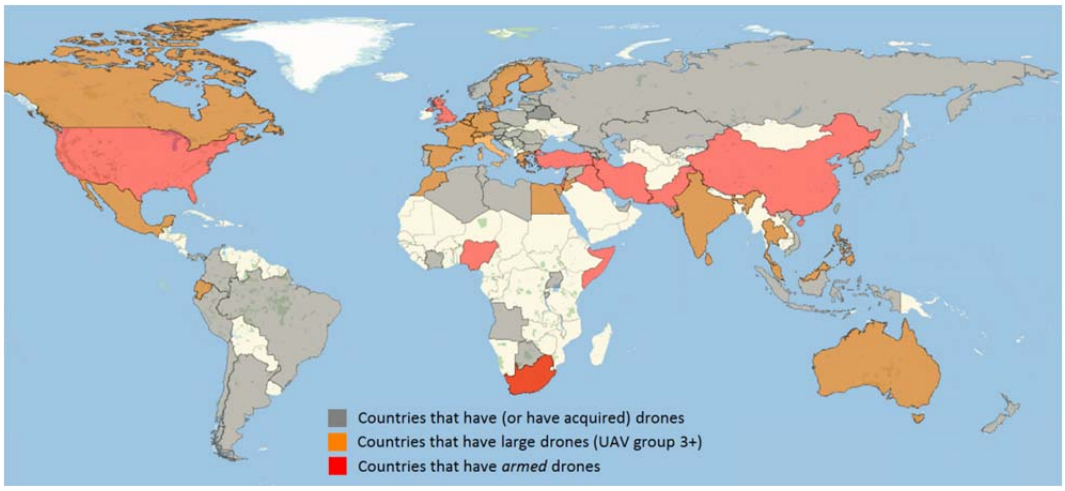
[58] H. Seck, "Marine Corps Shelves Futuristic Robo-Mule Due to Noise Concerns," *Military.com*, 22 Dec 2015: http://www.military.com/daily-news/2015/12/22/marine-corps-shelves-futuristic-robo-mule-due-to-noise-concerns.html?ESRC=todayinmil.sm.

[59] L. Greenemeier, "The Drone Wars: 9/11 Inspired Advances in Robotic Combat," *Live Science,* 3 Sep 2011: http://www.livescience.com/15908-drone-wars-september-11-anniversary.html.

(GAO) examined the worldwide proliferation of UAVs, noting that in the seven years since its last accounting, "the number of countries that acquired an unmanned aerial vehicle (UAV) system nearly doubled from about 40 to more than 75."[60]

Figure 3 provides a graphical depiction of the proliferation of UAVs, highlighted by three colors: (1) *grey*, to indicate countries that have either themselves developed drones or have acquired them elsewhere,[61] (2) *orange*, to indicate countries that have either class 2 or class 3 (i.e., "large") drones, and (3) *red*, to indicate countries that have *armed* drones. DoD classifies UAVs into five categories (Group 1 through Group 5), which are distinguished by maximum gross takeoff weight (MGTW), normal operating altitude (NOA), and maximum airspeed (see table 1).

Figure 3. Countries with unmanned aerial vehicles



References: *Jane's Unmanned Aerial Vehicles and Targets, 2011*; *US Unmanned Aerial Systems,* Congressional Research Service, 2012; *Information Sharing and End-Use Monitoring on Unmanned Aerial Vehicle Exports,* GAO Report, 12-536, Sep 2012; and J. Wilson, *2013 Worldwide UAV Roundup,* American Institute of Aeronautics and Astronautics, July–August 2013.

[60] *Agencies Could Improve Information Sharing and End-Use Monitoring on Unmanned Aerial Vehicle Exports*, U.S. Government Accountability Office Report GAO-12-536, Washington, DC, July 2012: http://www.gao.gov/assets/600/593131.pdf.

[61] Of the 31 counties that currently have large drones, 28 have either directly purchased some or all of them from another country or developed their drones with the help of another country. The main exporter of drone technology is Israel, which has exported its larger drone technology to 13 countries and assisted 4 others. France has directly exported to 3 countries, and the United States to 6, while helping at least one other country develop drone technology. Ref: C. Cole, "Is Drone Proliferation about to Explode?" *Drone Wars UK*, 25 May 2012.

Table 1. DoD UAV categories

Category	MGTW (lbs)	NOA	Airspeed (knts)	Category
Group 1	< 20	< 1200 above ground level (AGL)	< 100	*Raven*
Group 2	21 - 55	< 3500 AGL	< 250	*Scan Eagle*
Group 3	< 1320	<18,000 mean sea level (MSL)	< 250	RQ-7/ *Shadow*
Group 4	> 13320		Any	*RQ-1 / MQ-1 Predator*
Group 5	> 13320	> 18,000 MSL	Any	RQ-4 / *Global Hawk*

Ref: U.S. Army Unmanned Aircraft Systems roadmap 2010-2035, Army UAS CoE Staff, U.S. Army UAS Center of Excellence (ATZQ-CDI-C).

It is estimated that 31 countries have large military drones (i.e., Group-3+; see table 1). And the *weaponized* drone club has recently grown to 11 nations, including the United States, the United Kingdom, China, Israel, Pakistan, Iran, Iraq, Nigeria, Somalia, and South Africa (two non-state organizations—Hamas and Hezbollah—are also on the list).[62] As of this writing, eight countries have actually *used* armed drones in combat: the United States, Israel, the United Kingdom, Pakistan, Iraq, Nigeria, Iran, and Turkey (along with one non-state actor, Hezbollah).[63]

In the United States alone, over 1 million civilian drones were sold in 2015, and on a global level multiple millions of drones, ranging from small toy drones to larger commercial models, are sold and purchased.[64] Given this general availability of commercial drones, there is the growing concern that terrorists and insurgents will use UAVs for their attacks.[65] Attempts to do so date back to at least 1994, when the Japanese apocalyptic cult *Aum Shinrikyo* attempted (but failed) to release the nerve

[62] *World of Drones: Military*, International Security Data Site, New America Foundation: http://securitydata.newamerica.net/world-drones.html.

[63] Ibid.

[64] W. Zwijnenburg, "Terrorist drone attacks are not a matter of *if* but *when*," *Newsweek,* 29 April 2016: http://www.newsweek.com/drones-isis-terrorist-attacks-453867

[65] T. Burgers and S. Romaniuk, "The Next Generation of Terror: Swarming, Flying Bomb Robots," *The National Interest*, 21 Dec 2016.

agent *sarin* using remote-controlled helicopters equipped with aerial spray systems.[66] The next attempt by a terrorist group to use a UAV (also unsuccessful) was by Osama bin Laden, when, in July 2001, he tried using remote-controlled airplanes to deliver an improvised explosive device (IED) attack on G8 Summit leaders, then meeting in Genoa, Italy. Numerous terrorist plots and missions subsequently followed.[67] It is also well documented that Hezbollah has repeatedly attempted to attack Israel using commercially available UAVs with explosives,[68] and ISIS has started using small drones packed with explosives as weapons.[69]

Of course, both Russia and the People's Republic of China have invested heavily in developing unmanned systems (in all domains: air, land, sea, and underwater), and both are poised to become major global exporters of such systems.

China

A recent DoD assessment of China's military and security developments estimates that "China plans to produce upwards of 41,800 land- and sea-based unmanned systems, worth about $10.5 billion, between 2014 and 2023."[70] According to the report, at least four Chinese drones (the Xianglong, Yilong, Sky Saber, and Lijian, all introduced in 2013) are designed to carry precision-strike weapons, and the Lijian is a stealth drone. Some elements of China's emerging fleet of drones were revealed at the Zhuhai 2016 Airshow.[71] These included Cloud Shadow (a semi-stealthy roughly the size of the MQ-9) and the CH-5 (which has a wingspan of 21 meters, can carry payloads of up to one ton, has a flight time of 60 hours and a range of 6,500 km, and can link with other drones). A technology demonstration showed several dozen

[66] The same group later (in 1995) "successfully" carried out a sarin attack on a Tokyo subway. Ref: D. Rassler, *Remotely Piloted Innovation: Terrorism, Drones and Supportive Technology*, United States Military Academy, Oct 2016.

[67] R. Bunker, *Terrorist and Insurgent Unmanned Aerial Vehicles: Use, Potentials, and Military Implications*, U.S. Army War College, Strategic Studies Institute, August 2015. Bunker examines 24 terrorist and insurgent (attempted) use of UAVs between 1994 and 2015.

[68] M. Hoenig, *Hezbollah and the Use of Drones as a Weapon of Terrorism,* Federation of American Scientists, 2014: https://fas.org/wp-content/uploads/2014/06/Hezbollah-Drones-Spring-2014.pdf.

[69] B. Watson, "The drones of ISIS," *Defense One*, 12 Jan 2017: http://www.defenseone.com/technology/2017/01/drones-isis/134542/.

[70] *Annual Report to Congress: Military and Security Developments Involving the People's Republic of China 2015*, Office of the Secretary of Defense, 7 April 2015: https://www.defense.gov/Portals/1/Documents/pubs/2015_China_Military_Power_Report.pdf.

[71] J. Lin and P. Singer, "China's new fleet of drones: airshow displays the future of Chinese warbots and swarms," *Popular Science*, 4 Nov 2016.

drones flying in swarm flight patterns, coordinated by an interdrone communication network.

And in a recently reported live-fire test, Chinese CH-4 drones (1,300-kg UAVs with a 345-kg payload and a 35-hour flight endurance at 4,000-meter altitude; comparable to the US MQ-9 Reaper) fired their missiles on command from pilots over 1,000 km away. Earlier Chinese-made drones were limited to direct line-of-site communications from a ground station.[72]

Russia

Russia's "wake-up call" on the importance of having unmanned systems arguably came during the 2008 Russo-Georgian War.[73] Georgian government forces used Israeli-made Hermes Elbit 450 surveillance drones to conduct reconnaissance flights over the conflict regions (thereby gaining a demonstrable advantage), but Russia's own drones came too late to provide real-time intelligence. (Russian Defense Minister Anatoly Serdyukov reportedly forgot to sign an order authorizing their use.) This left the fighter jets and bombers that were sent on reconnaissance missions to compensate, needlessly vulnerable to attack. Since then, Russia has shown a commitment to developing unmanned systems (spurred, partly by the diminishing size of its army);[74] however, their efforts to date have lagged their Western and East Asian counterparts in reach, distance, and strike capability.[75] It was only in 2012 that the Russian Defense Ministry formed a department to manage drone research and development.[76]

Having learned its lesson in 2008, Russia has deployed over 16 types of drones for reconnaissance, surveillance, and targeting during its invasion of Ukraine.[77] Russia's

[72] J. Lin and P. W. Singer, "Chinese drones make key breakthrough, firing on command by satellite," *Popular Science*, 8 June 2016.

[73] N. Clayton, "How Russia and Georgia's 'little war' started a drone arms race," *GlobalPost,* 23 Oct 2012: http://www.pri.org/stories/2012-10-23/how-russia-and-georgias-little-war-started-drone-arms-race.

[74] M. Galeotti, "Russia turns to drones and robots as army shrinks," *Blouinnews*, 15 Dec 2013.

[75] S. Bendett, "How Russia's Military Plans to Counter the Pentagon's Drone Swarms," *The National Interest*, 10 Jan 2017.

[76] "'This is not a computer game'—Putin on drone use in Russia," Sputnik News: The Voice of Russia, 28 Nov 2013: http://sputniknews.com/voiceofrussia/news/2013_11_28/This-is-not-a-computer-game-Putin-on-drone-use-in-Russia-4577/.

[77] P. Tucker, "Is Russia Beating the U.S. in the Drone Race?" *The Fiscal Times*, 29 Sep 2016.

state media has also recently announced that the T-14 Armata tank may be the world's first fully autonomous tank.[78]

DoD drone-related funding

In FY01, DoD invested approximately $667 million on UASs.[79] In FY12, total spending (including both development and procurement) had increased to $3.9 billion.[80] And, in the proposed FY17 budget, DoD has allocated approximately $4.457 billion for drones (DARPA's FY17 budget includes $301.5 million for research into drones, autonomy and robotics, compared to $283.9 million in FY16, and $233.2 million in FY14).[81] The Navy's portion of the FY17 budget, which includes the Marine Corps, includes a total of $1.74 billion for drone procurement, research, and construction projects, the largest portions of which are earmarked for the MQ-4C Triton ($465 million), X-47B strike drone that is being retooled for aerial refueling ($90.4 million, for research), and the procurement and development of underwater drones ($279 million). This can be compared to $2.14 billion in FY16, and $1.2 billion in FY15).[82] The Army's FY17 budget includes $4.7 million to study swarming (with an additional $2.7 million to develop a counter-drone technology); and the Air Force's FY17 budget includes $52 million to improve human-machine interaction and the autonomous capabilities of unmanned vehicles.[83]

To get a sense of the rapid increase in the number of drones in DoD's inventory, observe that from 1994—when the Predator (RQ-1) made its maiden flight—DoD's inventory of unmanned aircraft grew to 163 in 2003 (with only 5 major programs), and to close to 11,000 in 2013 (distributed over the 13 major programs that cover all five UAV "group" categories, and account for 40% of all aircraft); see figure 4.

The deeper story (which will segue our narrative into a discussion of AI, in general, and autonomy, in particular) may be gleaned from what—at first sight—appears to be an inconsistency among DoD recent funding and policy trends (see insert plots A and B at bottom right of figure 4). For example, while DoD expects to increase its use

[78] A. Zemlianichenko, "Armata Designers: This May be the First Unmanned Drone Tank," Sputnik News, 13 June 2016: https://sputniknews.com/russia/201506131023298755/

[79] J. Gertler, *U.S. Unmanned Aerial Systems*, Congressional Research Service, CRS Report for Congress, 3 Jan 2012.

[80] *Program Acquisition Costs by Weapon System*, Office of the Under Secretary of Defense (Comptroller)/CFO, February 2011.

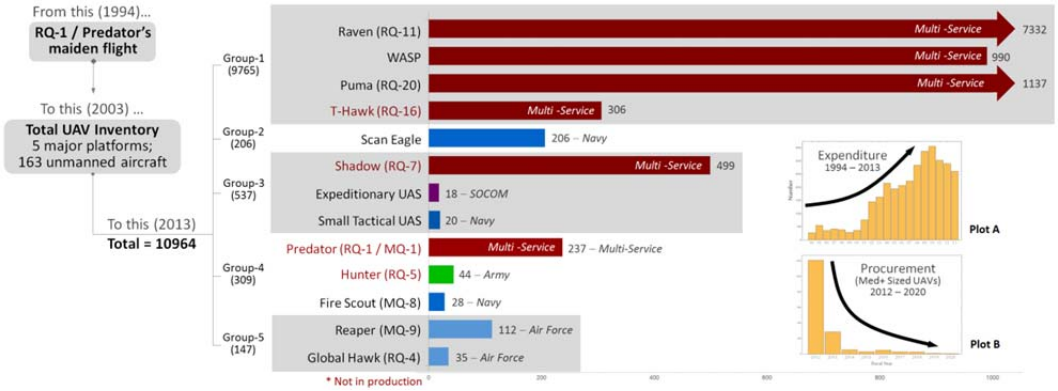
[81] D. Gettinger, *Drone Spending in the Fiscal Year 2017 Defense Budget*, Center for the Study of the Drone, Bard College, Feb 2016.

[82] Ibid., pp. 9-13.

[83] Ibid., pp. 3-8 and 14-17.

of unmanned systems by nearly 50% over the next several years—and total expenditure has risen (figure 4, Plot A)—funding for the procurement of new medium and larger UAVs (i.e., MQ-1, RQ-4, MQ-8, MQ-9, RQ-11) has dropped sharply (figure 4, Plot B): from 1,211 (in FY12) → 288 (in FY13) → 54 (in FY14) → … → 31 (in FY17).[84]

Figure 4. Inventory of major DoD UAVs



Ref: *Unmanned Systems Integrated Roadmap: FY2013-2038,* U.S. Department of Defense.

There are two plausible reasons for this:[85] (1) the thousands of UAVs purchased before 2012 were procured to address the high demands for intelligence and other ongoing missions in Iraq and Afghanistan (and there are sufficient numbers of current generation UAVs to meet the declining requirements as U.S. forces withdraw from those areas); and (2) the UAVs that have been procured thus far were primarily developed not as a result of DoD requirements but by commercial research and development.

DoD's recent budgets thus both reflect the reality of a transitioning commercial technology and portend the military's strategic shift to less permissive operating environments (i.e., the Asia-Pacific region). While the next generation of UAVs is likely to include a variety of "old" (albeit enhanced) technologies—thereby gaining greater speed, endurance, and stealth—and "new" technologies (not the least of which will be increasing levels of autonomy), DoD has not issued requirements or specifications, and thus appears "willing to let industry develop and put forward whatever that next generation will be."[86]

[84] DoD FY17 President's Budget Proposal, Release No: NR-046-16, 9 Feb 2016.

[85] J. Gertler, "How Many UAVs for DoD?" *CRS Insights*, 27 August 2015.

[86] Ibid.

This confluence of events represents an ample pool of general opportunities for military operations research analysis (MORA): (1) to help the DoD better understand emerging new AI- and robot-related technologies (e.g., by developing taxonomies and conceptual frameworks within which otherwise overly complex "systems" can be systematically examined); (2) to provide an impartial bridge between military requirements and the products of commercial and academic research (e.g., MORA organizations can be used as "go betweens" linking DoD and private technology industries);[87] and (3) to help develop new methods and analysis techniques, and new measures of effectiveness and performance, for studying unmanned systems with increasing levels of autonomy (beyond the "low hanging fruit" variety of assessing basic tradeoffs between manned and unmanned systems, calculating optimal paths, or developing targeting algorithms).[88]

Timeline of AI-, robot-, and swarm-related technologies

The idea of infusing inert objects with life and intelligence has existed for millennia, dating back at least to the ancient Greek myth of how Hephaestus, the god of fire, forged golden automaton-like statues to serve the gods.[89] The word "robot" is based on the Czech word *robota*, meaning "serf or slave," and first appeared in Karel Capek's 1921 play, *R.U.R.* ("Rossum's Universal Robots").[90] Then, in 1942, science fiction author Isaac Asimov published a short story called "Runaround," which introduced his well-known "Three Laws of Robotics."[91] After that, the word "robot" effectively became part of the common lexicon. The development of the first *physical*

[87] A recent example of a DoD-sponsored laison effort is the Defense Innovation Unit-Experimental (DIUx) office, established in 2015 in Sunnyvale, California, by Secretary of Defense Ashton Carter. DIUx's charter is to seek out innovative technologies and talent from Silicon Valley. Ref: M. Eaglen, "Tech-challenged Pentagon searches for a Silicon ally," American Enterprise Institute, 1 Feb 2016: https://www.aei.org/publication/tech-challenged-pentagon-searches-for-a-silicon-ally/.

[88] J. Cares and J. Dickman, editors, *Operations Research for Unmanned Systems*, Wiley, 2016.

[89] D. Gera, *Ancient Greek Ideas on Speech, Language, and Civilization*, Oxford University Press, 2003.

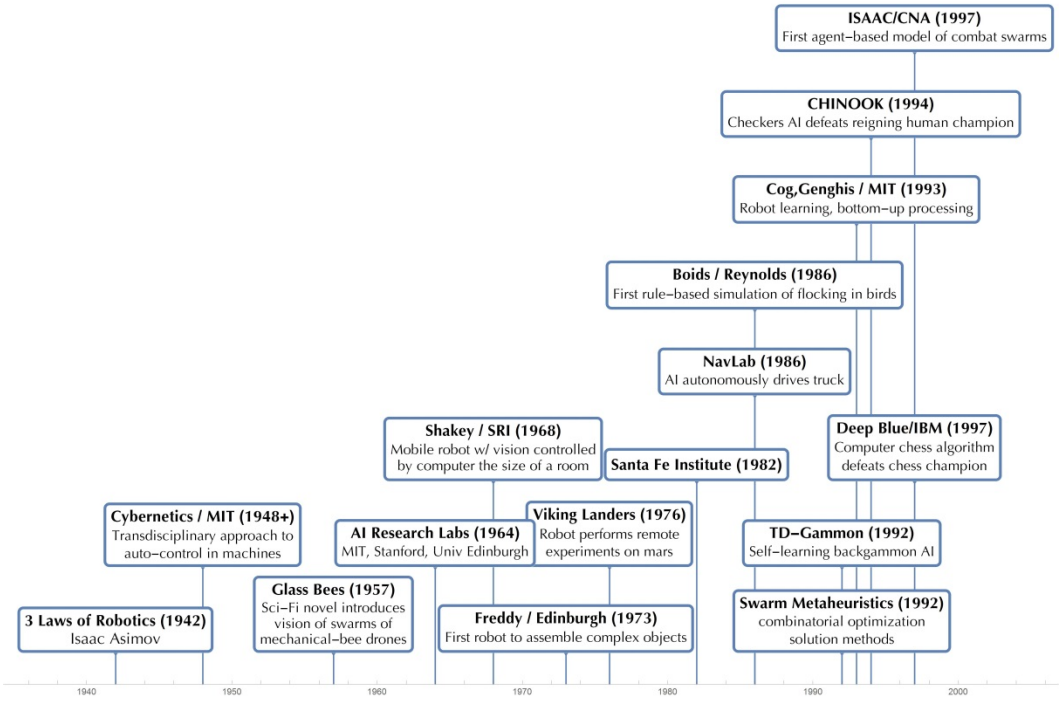
[90] In the beginning of the play, robots are synthetic humans that work in factories to produce low-cost goods. The play ends as these robots kill off the human race. Ref: K. Capek, *R.U.R.*, Penguin Classics, 2004.

[91] The "Three Laws of Robotic" are discussed in a later section of this report. "Runaround," in which the laws are quoted from an imaginary "Handbook of Robotics, 56th Edition, 2058 A.D.," appears in the short story collection: I. Asimov, *Robot Visions*, Roc, 1991.

instantiations of robots soon followed. The concept of a *swarm* was also introduced in a science fiction story—"The Glass Bees"—published by Ernst Jünger in 1957.[92]

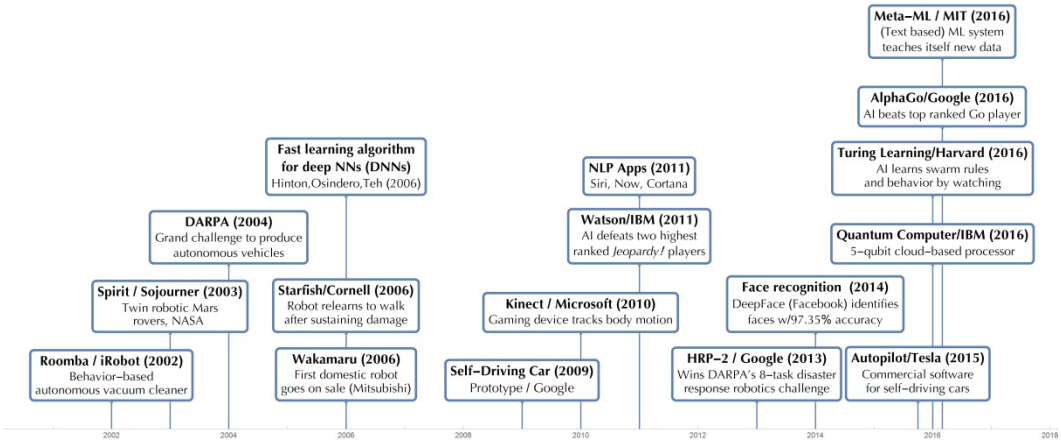
Figures 5 and 6 show timelines of selected milestones in the development of AI-, robot-, and swarming-related technologies. Figure 5 covers the years from 1942 to 1997; figure 6, the years from 2002 to 2016.

Figure 5. Timeline of selected milestones in the development of AI-, robot-, and swarming-related technologies (from 1942 to 1997)



[92] A major part of the story involves swarms of robotic bees (that are said to be much more efficient at gathering nectar than their real counterparts); the existence of humanoid robots is also alluded to. Ref: E. Juner, *The Glass Bees*, NYRB Classics, 2000.

Figure 6. Timeline of selected milestones in the development of AI-, robot-, and swarming-related technologies (from 2002 to 2016)



The key conceptual and technological development from which many aspects of modern AI and robotics research derives is *cybernetics*.[93] Inspired by and developed, in part, because of some specific military demands of World War II (e.g., the problem of automatic aiming and firing of anti-aircraft guns), cybernetics is a transdisciplinary approach to understanding *auto*-control in machines and other physical, biological, and social systems.[94] It is also the conceptual precursor of modern-day *complex systems theory* (discussed in a later section).[95] The word itself was introduced (along with the preliminary concepts describing its meaning) by Norbert Weiner, who was inspired by the Greek verb *kybernan*, which means "to steer, navigate, or govern."[96] Weiner defined cybernetics as "the scientific study of control and communication in the animal and the machine."[97] As a nascent discipline during World War II, cybernetics embodied three core concepts: *control*, *feedback*, and the *merging of human and machine*—concepts that, even without further explication (though discussions of each appear throughout this report), obviously overlap the main themes of this study.

[93] T. Rid, *Rise of the Machines: A Cybernetic History*, W. W. Norton & Company, 2016.

[94] A. Pickering, *The Cybernetic Brain*, University Of Chicago Press, 2011.

[95] G. Mobus and M. Kalton, *Principles of Systems Science*, Springer-Verlag, 2015.

[96] *Online Etymological Dictionary:* http://www.etymonline.com/index.php?term=govern.

[97] N. Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine*, MIT Press, 1948.

The timeline that unfolds after World War II consists of a series of (often "unexpected" [98] and "deceptively simple") nascent technologies that are, in hindsight, recognized as precursors of what later become major (and transformative) capabilities. The timeline also includes examples of some of the academic, commercial, and military research institutions that formalized and nurtured the study of AI, robotics, and/or swarms—e.g., DARPA,[99] which was founded in 1958[100] and offered the first grand challenge to produce an autonomous driverless car in 2004;[101] the first AI research labs established at MIT, Stanford University, and the University of Edinburgh in 1964; and the Santa Fe Institute,[102] founded in 1982, which pioneered the study of complex systems).

Other milestones include:

- The first rule-based simulation of swarm behavior in 1986 by computer scientist Craig Reynolds (which was called "Boids"[103] and lives on, in spirit if not detail, in many of today's robotic swarms).

- The first AI to beat the best human players at backgammon (TD-Gammon, in 1992), checkers (CHINOOK, in 1994), chess (Deep Blue, in 1997), and Go (AlphaGo, in 2016). The latter accomplishment is particularly noteworthy: in 20 years, the state-of-the-art in AI had progressed from being able to defeat the (then) world champion in chess—already a laudable accomplishment—to beating one of the world's top-ranked players of a game, Go, that is so much more "complex" (as a problem for an AI to "solve")

[98] In deference to available space, we leave out of the narrative several "unexpected" disruptive technologies that have had a major (albeit more implicit) impact on the subject domains of this report (e.g., the development of the transistor and integrated circuits, the internet, stealth, and GPS). Ironically, it is AI's potential to become a major disruptive technology—not just for the military, but for the world in general—that directly inspired this study. A monograph that explores ways of describing, and forecasting, disruptive technologies was recently published by the National Research Council: *Persistent Forecasting of Disruptive Technologies*, Committee on Forecasting Future Disruptive Technologies, National Academies Press, 2010.

[99] http://www.darpa.mil/.

[100] A. Jacobsen, *The Pentagon's Brain: An Uncensored History of DARPA, America's Top-Secret Military Research Agency*, Back Bay Books, 2016.

[101] J. Hooper, "From DARPA Grand Challenge 2004: DARPA's Debacle in the Desert," *Popular Science*, June 2004.

[102] https://www.santafe.edu/.

[103] C. Reynolds, "Flocks, herds, and schools: A distributed behavioral model," *Computer Graphics* 21, no. 4, 1987.

that priori to AlphaGo's victory, most AI experts believed that it could not be done for another 15–20 years.[104]

- The first prototype self-driving car developed by Google in 2009,[105] with Tesla's *Autopilot*[106] now commercially available and installed in the vehicles it sells.

- AI software that is able to understand natural language (in spoken form) and to recognize faces in images well enough to be commercially viable (e.g., Apple's *Siri*, Google's *Now*, and Microsoft's *Cortana* apps, respectively;[107] and Facebook's *DeepFace* algorithm,[108] that reportedly identifies faces with a 97.35% accuracy).

- In 2016, MIT researchers introduced an AI system that effectively uses a meta-learning algorithm that allows it to learn to extract text information *on its own* (when traditional "training" data are not available or are scarce).[109]

- A potentially revolutionary AI machine-learning method—called *Turing Learning*,[110] and developed at Harvard in 2016—that includes the first-ever demonstration of an AI system that can automatically infer the behavior of physical robot swarms simply by watching.

Other recent milestones (which do not appear in figure 5 and 6) include:

[104] http://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/.

[105] https://www.waymo.com/journey/.

[106] https://www.tesla.com/autopilot.

[107] J. Dunn, "We put Siri, Alexa, Google Assistant, and Cortana through a marathon of tests to see who's winning the virtual assistant race — here's what we found," *Business Insider*, 4 Nov 2016.

[108] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *Conference on Computer Vision and Pattern Recognition* (CVPR), 24 June 2014: https://research.fb.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/.

[109] K. Narasimhan, A. Yala, and R. Barzilay, "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning," presented at EMNLP 2016, arXiv:1603.07954v3: https://arxiv.org/abs/1603.07954.

[110] W. Li, M. Gauci, and R. Gross, "Turing learning: a metric-free approach to inferring behavior and its application to swarms," *Swarm Intelligence* 10, no. 3, September 2016: http://link.springer.com/article/10.1007%2Fs11721-016-0126-1.

- Engineers from the University of Toronto, Canada, have recently introduced a novel approach to train neural networks (called "heuristic training") that relies not on labelled data curated by human programmers but rather on natural-language input consisting of high-level assertions that describe basic properties of a system that a neural net will then "train itself" to identify patterns in. An early prototype has already outperformed conventional machine learning algorithms by 160 percent.[111]

- Google's on-line translation algorithm was recently enhanced with: (1) a neural network that trains on entire sentences at once (neural net "training" is discussed in a later section), thus providing it with a deeper context in which to come up with a translation, and (2) an ability to simultaneously translate among multiple pairs of languages (and just between a single pair of languages, as is common practice).[112] A surprising result is that Google's enhanced translation algorithm can now translate between two languages that it has *not been directly trained on*. The researchers responsible for this work suspect that the AI system has effectively invented its own "interlingua" language—i.e., an intermediary semantic core in which sentences with the same meaning are represented in similar ways independent of the specific languages (albeit one that resides entirely "internally" and is not understandable or usable by humans).

- Researchers from *Google Brain*, Google's deep learning project, have recently unveiled an AI "secure communications" system that can effectively *invent its own encryption scheme*, and without being taught specific cryptographic algorithms.[113] Importantly, the researchers themselves do not know exactly how the encryption method works, since neural-network-based techniques such as the one used for this study do not easily reveal how their "solutions" actually work.

- Google's *DeepMind* has been enhanced with an ability to *interact with its environment* (via virtual actuators) in order to solve problems.[114] The testbed

[111] W. Guo and P. Aarabi, "Hair Segmentation Using Heuristically-Trained Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems* 99, 2016.

[112] Q. Le and M. Schuster, "A Neural Network for Machine Translation, at Production Scale," Google Research Blog, 27 Sep 2016: https://research.googleblog.com/2016/09/a-neural-network-for-machine.html.

[113] M. Abadi and D. Andersen, "Learning to Protect Communications with Adversarial Neural Cryptography," arXiv:1610.06918v1: https://arxiv.org/abs/1610.06918.

[114] M. Denil, P. Agrawal, T. Kulkarni, et al., "Learning to perform physics experiments via deep reinforcement learning," under review as a conference paper to ICLR 2017: https://arxiv.org/pdf/1611.01843v1.pdf.

tasks were to determine the number and the heaviest of five blocks stacked in a tower. Apart from being rewarded when it answered correctly, and penalized when it gave an incorrect answer, no other "instructions" were provided to the AI. The system effectively "solved" the problem on its own by manipulating its physical environment. This "toy-level" demonstration is a key first step toward developing AI systems that learn to develop a common sense understanding of our physical world on their own.

"Third Offset Strategy"

DoD's near– to mid–term interest in AI, robotics, autonomous systems, and other innovative technologies will inevitably be shaped by then defense secretary Chuck Hagel's announcement of a new Defense Innovation Initiative—which included the Third Offset Strategy (TOS)—on 15 November 2014.[115] An *offset strategy* (OS) is a general set of peacetime competitive policies designed to generate and sustain a strategic advantage over one's main adversaries. It is less about formulating new (or adapting old) theories of warfare, and more about brainstorming what ought to be done to gain and/or strengthen the military's competitive edge. Judging by the only two existing offset strategies prior to TOS (see below), an OS is not myopically focused on technology alone; rather, its goal is to articulate a mix of technology and shifts in operational perspective that is judged best able to achieve the requisite strategic advantage.[116]

For example, the *First Offset* was part of President Eisenhower's "New Look" in the 1950s at the start of the Cold War, and introduced tactical nuclear weapons to match Soviet numerical and geographical advantages along German border.[117] The perceived imbalance at the time was the Soviet Union's geographical advantage over the United States in Western Europe, so a strategy was born to regain the advantage by exploiting the superiority of our nuclear weapons technology. Some key investments that were fueled by the First Offset included much smaller-sized nuclear weapons, the development of intercontinental ballistic missiles (ICBMs), expanded aerial

[115] C. Hagel, Transcript of Keynote speech delivered at Reagan National Defense Forum Keynote, Ronald Reagan Presidential Library, Simi Valley, CA, Nov. 15, 2014: http://www.defense.gov/News/Speeches/Speech-View/Article/606635.

[116] K. Lange, "3rd Offset Strategy 101: What It Is, What the Tech Focuses Are," DoD Live, 30 March 2016: http://www.dodlive.mil/index.php/2016/03/3rd-offset-strategy-101-what-it-is-what-the-tech-focuses-are/.

[117] T. Walton, "Securing the Third Offset Strategy," *Joint forces Quarterly*, National Defense University, Issue 82, 3rd Quarter 2016.

refueling, and enhanced air/missile defense networks. Notably, most major First-Offset-related breakthrough technologies were funded by the Department of Defense.

Harold Brown's *Second Offset* was spurred on after the Soviet Union's nuclear weapon technology and delivery systems essentially caught up to those of the United States, and strategic thinking turned to regaining a non-nuclear tactical advantage.[118] Key investments that resulted from the Second Offset included new intelligence, surveillance, and reconnaissance (ISR) platforms and battle management capabilities, precision-strike weapons, stealth aircraft, smart weapons and sensors, and the burgeoning tactical exploitation of space (e.g., GPS). The key technological drivers in this case were digital microelectronics and information technology.[119]

And so we come to the TOS, which was announced formally on 15 November 2014, but the rudiments of which were openly discussed earlier.[120] The perceived "imbalance" for this third go-around is a combination of two factors: *shrinking force structure* and *declining technological superiority.*[121] The goal is not the acquisition of next-generation technologies, per se, but a combined re-evaluation of technological innovations and new concepts of operations.[122]

In terms of a more detailed breakdown of what TOS portends for specific investments, five core building blocks were outlined by Deputy Secretary of Defense Robert Work during a keynote address at the CNAS Inaugural National Security Forum held on 14 December 2015:[123]

[118] H. Brown, Secretary of Defense, *DoD Annual Report, Fiscal Year 1982*: http://www.dtic.mil/dtic/tr/fulltext/u2/a096066.pdf.

[119] J. McGrath, "Twenty-First Century Information Warfare and the Third Offset Strategy," *Joint forces Quarterly*, National Defense University, Issue 82, 3rd Quarter 2016.

[120] C. Hagel, Transcript of Keynote speech delivered at the *Southeastern New England Defense Industry Alliance*, Newport, Rhode Island, Sept. 3, 2014: http://www.defense.gov/News/Speeches/Speech-View/Article/605602.

[121] M Eaglan, "What is the Third Offset Strategy?" *RealClear Defense*, 16 Feb 2016: http://www.realcleardefense.com/articles/2016/02/16/what_is_the_third_offset_strategy_109034.html.

[122] A. Carter, Secretary of Defense, speech delivered at the "The Path to an Innovative Future for Defense," CSIS Third Offset Strategy Conference, 28 October 2016: http://www.defense.gov/News/Speeches/Speech-View/Article/990315/remarks-on-the-path-to-an-innovative-future-for-defense-csis-third-offset-strat.

[123] Keynote by Deputy Secretary of Defense Robert Work at the CNAS Inaugural National Security Forum, December 14, 2015: http://www.defense.gov/News/Speeches/Speech-View/Article/634214/cnas-defense-forum.

- *Autonomous Deep-Learning Systems*
  - Examples: The Air Force Research Laboratory's (AFRL's) Autonomous Defensive Cyber Operations (ADCO); National Geospatial Agency's (NGA's) Coherence Out of Chaos program (deep-learning-based queuing of satellite data for human analysts); Israel's Iron Dome air defense system
- *Human-Machine Collaborative Decision-Making*
  - Examples: 2005 human-chess collaboration that defeated field of chess champions and grandmasters; F-35 helmet portrayal of 360 degrees on heads-up display
- *Assisted Human Operations*
  - Examples: wearable electronics, heads-up displays, exoskeletons
- *Advanced Manned-Unmanned System Operations*
  - Examples: Army's Apache and Gray Eagle UAV,[124] and Navy's P-8 aircraft and Triton UAV; small service vessels operating as swarms, with one mission commander simultaneously directing the swarm itself
- *Network-Enable, Semi-Autonomous Weapons Hardened To Operate in a Future Cyber/EW Environment*
  - Example: modified Air Force's Small Diameter Bomb (SDB) for operating in GPS-denied environment

While each of these building blocks obviously describes long-term commitments—an assessment of the requisite operational analysis of which is the central focus of this white paper—a number of short-term investments consistent with their overarching vision have already made their way into DoD's FY17 budget. Of the $71.8 billion allocated for research and development (R&D)—4% greater than in the budget for FY16—$12.5 billion is budgeted for science and technology (S&T), and $18 billion is spread over the Future Years Defense Program's (FYDP's) five year plan. The latter includes $3 billion for human-machine collaboration and teaming, $1.7 billion for cyber and EW issues, and $500 million for expanding war gaming and operational concept tests and demonstrations.[125] While specific TOS technology represents only $35 million of the overall S&T budget, $902 million is allocated for the Strategic

[124] http://www.ga-asi.com/gray-eagle.

[125] A. Mehta, "Defense Department Budget: $18B Over FYDP for Third Offset," *Defense News*, 9 Feb 2016: http://www.defensenews.com/story/defense/policy-budget/budget/2016/02/09/third-offset-fy17-budget-pentagon-budget/80072048/.

Capabilities Office (SCO)—which, it has been argued, is DoD's way of effectively "jumpstarting" TOS's long-term vision by focusing on improving TOS-like current-generation technologies.[126] Finally, $45 million is allocated to the *Defense Innovation Unit-Experimental* (DIUx) office, established in 2015 in Sunnyvale, California, by Secretary of Defense Carter to seek out innovative technologies and talent from Silicon Valley.[127]

TOS-specific thrusts in the FY17 budget include:[128]

- *Anti-access and area denial (A2/AD):* continued development of Air Force and Navy aviation propulsion development programs, new counter-space, and a Navy autonomous cargo re-supply platform
- *Guided munitions:* a program to counter hardened and deeply buried targets, experimentation with hypersonic weapons, and development of alternative guidance technologies to reduce reliance on GPS
- *Undersea warfare:* acceleration of traditional investments focusing on quieting and sensing improvements, $200 million portfolio of a diverse set of unmanned undersea vehicles
- *Cyber and electronic warfare:* $49 million allocated to accelerated development of the Advanced Anti-Radiation Guided Missile next-generation anti-radar missile, and cross-service investment in aircraft countermeasures against electronic warfare
- *Human-machine teaming:* accelerated development of the machine-aided Joint Precision Approach Landing System, a new program for a more intelligent logistics system and several unmanned systems projects
- *War gaming and concepts development:* $21 million allocated to the Navy's Fleet Experimentation program, and $60 million for naval rapid acquisition programs (such as Rapid Prototype Development and Unmanned Rapid Prototype Development)

[126] S. Freedberg, Jr., "Strategic Capabilities Office Is 'Buying Time' For Offset: William Roper," *Breaking Defense*, 18 July 2016: http://breakingdefense.com/2016/07/strategic-capabilities-office-is-buying-time-william-roper/.

[127] M. Eaglen, "Tech-challenged Pentagon searches for a Silicon ally," *American Enterprise Institute*, 1 Feb 2016: https://www.aei.org/publication/tech-challenged-pentagon-searches-for-a-silicon-ally/.

[128] M Eaglan, "What is the Third Offset Strategy?" *RealClear Defense*, 16 Feb 2016.

Accelerating technological change

The increasingly rapid pace of technological change, particularly in the fields of AI and robotics, is well known and documented.[129] For example, Ray Kurzweil (who directs *Google*'s research on machine intelligence and natural language understanding), has argued that technology, like biology, is an evolutionary process whereby the information-processing tools and methods of prior generations are used to generate those of the next. As improvements accrue and evolve, the time between successive advancements in order and capability decreases exponentially.[130] Moreover, according to Kurzweil's "Law of Accelerating Returns," if a technology stalls or comes up against some form of barrier impeding further progress, a new technology will be invented to militate the presence of the barrier.

Figure 7 shows, as a microcosm of a much larger space of general engineering and technology innovations,[131] a timeline of the accelerating growth of *computer power*, measured in terms of raw computations per second (CPS).[132] Note that the CPS curve is logarithmic, meaning that each unit increment along the ordinate on the plot represents a jump of 10× the prior value. Sequenced from the bottom to the top of the figure, the four horizontal dashed lines represent the approximate CPS values for an insect's brain, a mouse's brain, a human's brain, and all human brains on the planet, respectively.

The vertical dashed line, the bottom of which is buttressed against the label "Now," is centered on the year 2016 (i.e., this report's release date). The alternating shades of gray, from left to right, represent overlapping technological epochs, and range from an era of *electromechanical devices*, to *solid-state relays*, *vacuum tubes*, *transistors*, and *integrated circuits*. In addition, six AI-related milestones are highlighted in red along the bottom of the CPS timeline curve (all are discussed later in this report): neural nets (introduced in 1943); reinforcement learning (a technique that is, in part, responsible for Google's *AlphaGo*'s recent defeat of Lee Sedol in Go, and introduced in 1973); the backpropagation algorithm (that allows neural-nets to "learn"), introduced in 1986; IBM Deep Blue's landmark victory over world champion Gary

[129] *Timeline of Computer History:* http://www.computerhistory.org/timeline/ai-robotics/; J. Goodell, "Inside the Artificial Intelligence Revolution: A Special Report: Parts 1 & 2," *Rolling Stone Magazine*, February & March 2016.

[130] R. Kurweil, *The Singularity is Near*, Viking Press, 2005.

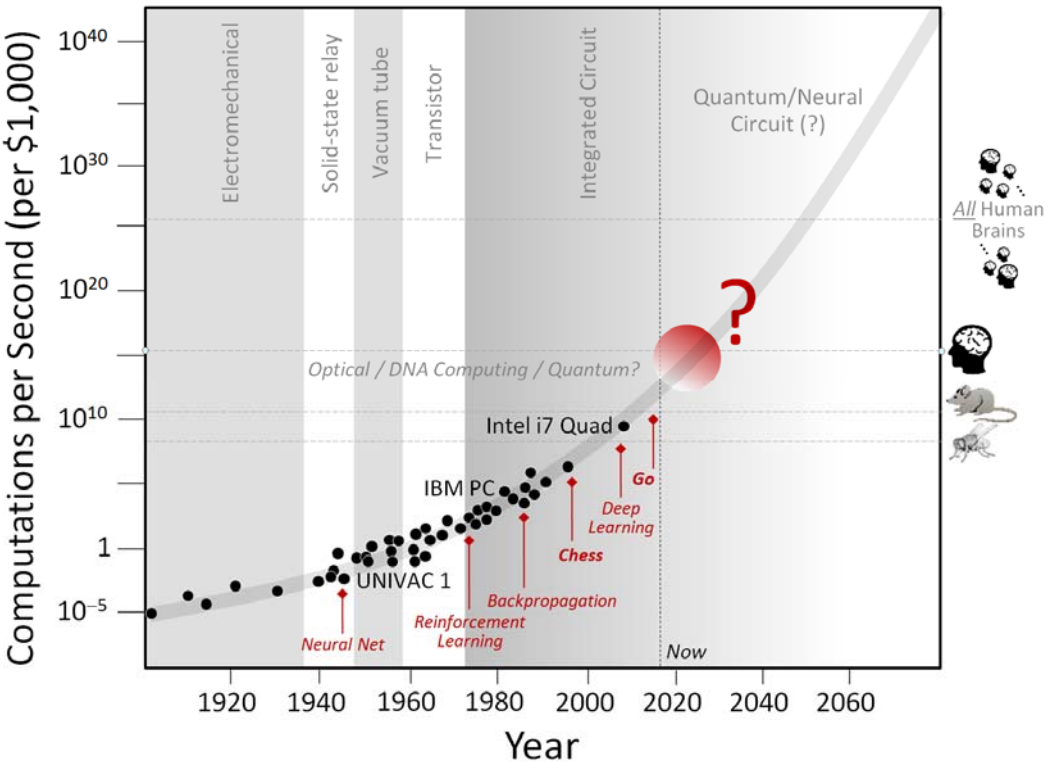
[131] See, for example: T. Jackson, editor, *Engineering: An Illustrated History from Ancient Craft to Modern Technology*, Hselter Harbor Press, 2016.

[132] C. Moore and A. Mertens, *The Nature of Computation*, Oxford University Press, 2011.

Kasparov in chess, developed in 1997; the "deep learning" algorithm (used by *AlphaGo* and other recent AI systems), published in 2006; and *AlphaGo*'s historic win over the reigning world champion human player in Go in 2016.

The red disk that appears just to the right of center of the figure denotes the area of uncertainty of the expected continued growth in CPS in the coming decade.

Figure 7. Accelerating growth of computing power



Ref: https://upload.wikimedia.org/wikipedia/commons/d/df/PPTExponentialGrowthof_Computing.jpg

The key takeaway from figure 7 is the observation that—as of this writing (November 2016)—the exponential CPS-vs-time curve is *about a decade away* from crossing the line that denotes the raw computational power of a human brain (the value of which is estimated to be between 10^{15} and 10^{16} CPS).[133] The precise value does not matter; nor does a "one human brain equivalent" of CPS represent a special barrier (such as the "speed of sound" for a jet) at which something magical happens. However, it does denote a computational threshold vastly beyond what our experience with computational technology has thus far prepares us for.

[133] N. Bostrom, "How long before superintelligence?" *Int. Jour. of Future Studies* 2, 1998.

Of course, no one can predict with any certainty when (or even if) the "one human brain equivalent" of CPS threshold will be achieved, or what impact this will have on business, culture, the military, or the world at large. But if recent advances offer even the most modest of clues, it is that we will be *surprised*. *AlphaGo*'s victory over Lee Sedol in 2016 took place less than a decade after IBM's *DeepBlue* defeated Gary Kasparov; a victory that, right up until it happened, most AI researchers believed was still decades in the future (and some wondered whether an AI system could ever be taught to play well enough to defeat a top-ranked human player). This is because the innate complexity of Go, as a game (as measured by, say, the number of "moves" an AI system has to search through before deciding on a move), exceeds that of chess by almost 240 orders of magnitude.[134] Yet, the advances in computation and machine-learning techniques made in just nine years were sufficient to "achieve the (recently-believed-to-be) unachievable."

Just a year before *AlphaGo*'s 2016 victory, essentially the same technique was used, for the first time, to master a diverse range of Atari 2600 games to a superhuman level with only the raw pixels and scores as inputs.[135]

In less than a year after *AlphaGo*'s landmark achievement, its underlying machine-learning technology has achieved several additional milestones, including *navigation* (in which an external memory is combined with deep learning—discussed later in the report—to build an AI that uses basic reasoning to "self learn" how to navigate the London underground),[136] and *encryption* (in which neural networks "teach themselves" how to encrypt messages, without relying on any a priori cryptographic algorithms).[137] Moreover, AlphaGo has continued to improve on its already core accomplishment: in January 2017, *DeepMind* founder Demis Hassabis revealed in a tweet that a theretofore anonymous online player known only as "Master"—who had been regularly beating the world's best Go players, including the world's top-ranked player 50 out of 51 games (drawing the one game it did not win)—was, in fact, an updated version of AlphaGo.[138]

[134] While chess is played on an 8-by-8 board, tournament-level Go is played on a 19-by-19 board, albeit with effectively two pieces. J. Burmeister, "The challenge of Go as a domain for AI research: a comparison between Go and chess," *Intelligent Information Systems*, 1995.

[135] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature* 518, 26 February 2015.

[136] A. Graves et al., "Hybrid computing using a neural network with dynamic external memory," *Nature* 538, 27 October 2016.

[137] M. Abadi, D. Andersen, "Learning to Protect Communications with Adversarial Neural Cryptograph," 21 Oct 2016: https://arxiv.org/abs/1610.06918v1.

[138] T. Revell, "DeepMind's AlphaGo is secretly beating human players online," *New Scientist*, 4 Jan 2017.

Evolution of DoD's interest in autonomy

> We don't have to develop new planes. We don't have to develop fundamentally new weapons. But we have to work the integration and the concept of operation. And then you have a completely new capability, but you don't have to wait long at all.
>
> —William Roper (Director, Strategic Capabilities Office, Office of the Secretary of Defense

DoD's interest in autonomy as a distinct operational capability goes back only a relatively few years to 2011, when it appears as the sixth (out of seven) S&T investment priority for FY12-17 planning in a memorandum released by the secretary of defense.[139] Prior to this memo, autonomy was used mainly as an ill-defined "catch-all" term to label general future technology enhancements of unmanned platforms, and was not seen as a specific system ability to perceive, reason, or plan.[140]

The Army's 2009 *Robotics Strategic White Paper*[141] was among the first by a Service to formally define a robot: "a man-made device capable of sensing, comprehending, and interacting with its environment." But this "pioneering" document refers only to innate hardware characteristics, and does not include any capabilities for formulating plans and making dynamic decisions. DoD's first *Unmanned System Integrated Roadmap* (USIR),[142] also released in 2009, contains over 80 references to autonomy and autonomous systems, and devotes an entire section to outlining how autonomy can be incorporated into future operations. The USIR represents the first systematic focus on autonomy within DoD, identifying it as the second of seven challenges facing all military Services. The USIR also lists Manned-Unmanned (MUM) Teaming (MUM-T) as a related challenge, and introduces the issue of "trust" as an element of autonomy that will require the development of new verification and validation (V&V) and testing and evaluation (T&E) techniques.

---

[139] Science and Technology (S&T) Priorities for Fiscal Years 2013-2017 Planning, Memorandum, Secretary of Defense, 19 April 2011.

[140] B. Grabowski, *Anticipating the Onset of Autonomy: A Survey of the DoD*, Armed Service, and other Federal Agencies' Outlook on Autonomy, MITRE Technical Report MP130118, 2013.

[141] *Robotics Strategy White Paper*, Army Capabilities Integration Center – Tank-Automotive Research and Development, Engineering Center Robotics Initiative, 19 March 2009.

[142] *Unmanned Systems Integrated Roadmap: FY2011-2036*, U.S. Department of Defense: http://www.acq.osd.mil/sts/docs/Unmanned%20Systems%20Integrated%20Roadmap%20FY2011-2036.pdf.

In 2010, the National Aeronautics and Space Administration (NASA) released its *Robotics, Tele-Robotics and Autonomous Systems Roadmap*,[143] in which autonomy "in the context of a system (robotic, spacecraft, or aircraft)" is defined as "…the capability for the system to operate independently from external control," and an autonomous system is defined as a system that "resolves choices on its own," though in the context of working towards goals "provided by another entity." Specific attributes include the ability for complex decision-making, including autonomous mission execution and planning, the ability to understand system state and react accordingly, and the ability to self-adapt in changing environments.

DoD's Defense Science Board's (DSB's) 2012 report on autonomy (DSB/2012)[144] is the first defense-sponsored study to discuss autonomy not as an innate system property, but as a *capability* to perform a set of functions while coupled to a dynamic environment: "Autonomy is better understood as a capability (or a set of capabilities) that enables the larger human-machine system to accomplish a given mission, rather than as a 'black box' that can be discussed separately from the vehicle and the mission."[145] Above all, human-machine teaming is stressed: "all autonomous systems are joint human-machine cognitive systems."

The report disentangles autonomy from the system platform by shifting the focus from hardware to software, thereby effectively recasting the development of autonomy (as viewed by earlier DoD and Service-centric studies) as being "primarily a software endeavor, which is a shift from traditional hardware oriented, vehicle centric development." The general consequences of this shift—which, as far as the author of this report is aware, remains unchallenged by any subsequent defense-sponsored reports, including DSB's more recent 2016 study (see discussion below)—and the impact it is likely to have on DoD's acquisition process, are potentially far-reaching. DSB/2012 emphasizes that software development generally lies "outside of the current hardware-oriented, vehicle-centric development and acquisition processes. Program managers may not know how to specify autonomy software, developers may not have sufficient expertise to write autonomy software, and testing and evaluation has few metrics and test beds for verification and validation." Autonomy's unique acquisition challenges are examined in a later section.

[143] *Robotics, Tele-Robotics, and Autonomous Systems Roadmap*, National Aeronautics and Space Administration, Nov 2010: http://www.nasa.gov/pdf/501622main_TA04-Robotics-DRAFT-Nov2010-A.pdf.

[144] *The Role of Autonomy in DoD Systems*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012: http://www.acq.osd.mil/dsb/reports/DSBSS15.pdf.

[145] Ibid., p. 21.

An important part of DSB/2012's shift from platform to capability consists of effectively doing away with the (then fashionable, though myopic) unidimensional views of autonomy—whereby a system's autonomy is assigned a numerical 'rank' on a sliding scale—and replacing them with an as-yet-undefined multidimensional conceptual framework (an idea that will be revisited in a later section):

> Treating the levels of autonomy as a developmental roadmap misses the need to match capabilities with the dynamic needs of the task or mission and directs programming attention away from critical, but implicit, functions needed for overall system resilience and human trust in the system. The mismatch of capabilities leads to gaps in functionality that have to be filled with additional manpower, creates vulnerabilities when unforeseen conditions arise and prevents rapid adaption or re-tasking of unmanned systems for new missions. Programming attention to the machine often means a lack of focus on the interfaces and tools that confirm to the operators and commanders that the system is performing mission priorities; without these interfaces and tools, there is no trust in the overall system.[146]

Figure 8 shows a timeline (2012-2016) of *unmanned-systems*– and *autonomy*–related DoD directives, memos, and DSB reports that have appeared since the DSB/2012 report. While we obviously do not have the space in this paper to discuss these reports, it is nonetheless instructive to summarize the evolution in thinking that this group represents as a whole.[147]

- **(July 2012) Defense Science Board Task Force Report (TF): *The Role of Autonomy in DoD Systems*[148]**
  - Shifts from systems and platforms to capabilities and decisions.
  - Advocates moving away from single-valued "ranks" of autonomy to developing a multidimensional conceptual framework.

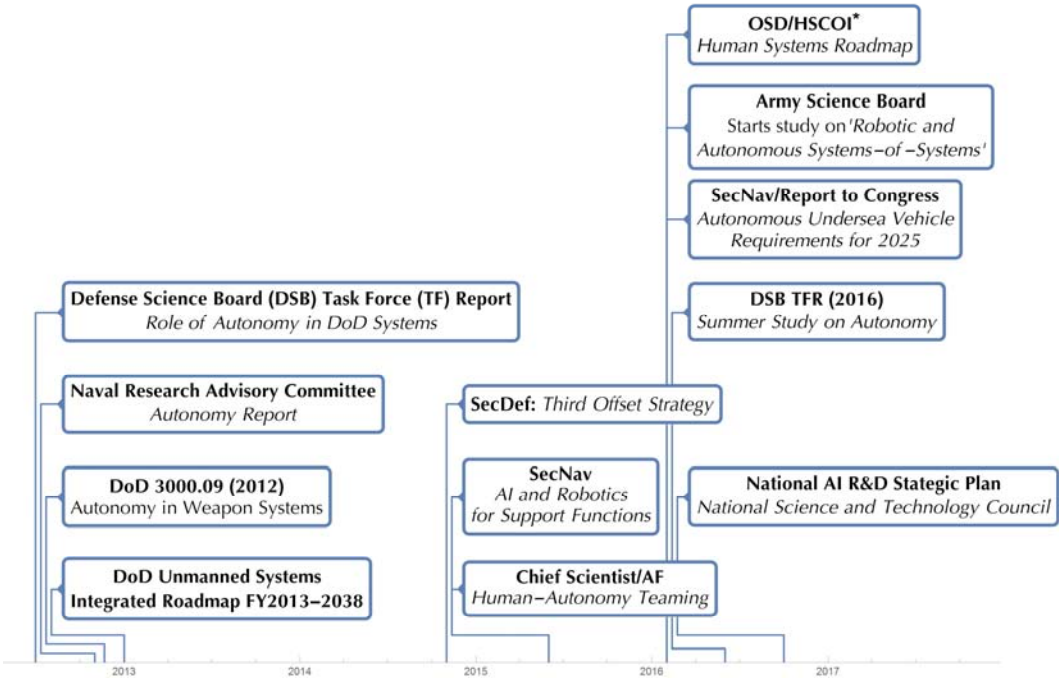
[146] Ibid., p. 24.

[147] In addition to the individual references themselves, we also follow: J. Caton, *Autonomous Weapon Systems: A Brief Survey of Developmental, Operational, Legal, and Ethical Issues*, U.S. Army War College, Strategic Studies Institute, Carlisle, PA, Dec 2015, and B. Grabowski, *Big Picture for Autonomy Research in DoD*, Keynote presentation at the Safe and Secure Systems and Software Symposium, held 09-11 June 2015, Dayton, Ohio, Air Force Research Laboratory: http://www.mys5.org/Proceedings/2015/Day_1/2015-S5-Day1_0805_KEYNOTE_Grabowski.pdf.

[148] *The Role of Autonomy in DoD Systems*, DoD, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

– Lack of effective coordination of research and development (R&D) efforts across military services.
– Autonomy entails unique acquisition challenges.
– Operational challenges are created by the urgent deployment of unmanned systems to theater without adequate resources or time to refine concepts of operations and training.
– States: "While currently fielded unmanned systems are making positive contributions across DoD operations, autonomy technology is being underutilized as a result of material obstacles within the Department that are inhibiting the broad acceptance of autonomy and its ability to more fully realize the benefits of unmanned systems."

Figure 8. Recent timeline (2012-2016) of directives, memos, and reports related to unmanned-systems and autonomy



* HSCOI (Human Systems Community of Interest) consists of senior officials from the U.S. Army, Navy, Marine Corps, Air Force, and DARPA. It is overseen by Assistant Secretary of Defense for R&E and Assistant Secretary of Defense for Health Affairs.

- **(Oct 2012) Naval Research Advisory Committee Report: *How Autonomy can Transform Naval Operations***[149]
  - Finds that autonomy represents a transformational, and potentially disruptive, capability; currently still driven by technology "push."
  - Identifies two essential keys to implementation of autonomy as a transformational capability: building a *community* and building *trust.*
  - Trust-building begins in the design and development phases by requiring Fleet involvement throughout the development process (not just in the final experimentation stage).
  - Potential near-term applications include: ocean monitoring, ISR, MCM, force protection, hull maintenance, and logistics).
  - Long-term opportunities include: the capacity to operate in A2/AD environment, mine clearing, antisubmarine warfare (ASW), and *in situ* ISR data processing (to reduce analyst load).
  - Recommendations include: establishing a Naval autonomy community (composed of technical, acquisition, requirements, and operational experts to focus on autonomy for Naval needs); periodically assessing global autonomy markets that may be relevant to its efforts; and developing protocols to support autonomous systems testing and "trust building."
- **(Nov 2012) DoD Directive (DoDD) 3000.09, *Autonomy in Weapon Systems***[150]
  - Establishes policy for development and use of autonomous systems
  - Shows awareness of unique challenge of autonomy-related testing and V&V
  - Puts emphasis and constraints on autonomous weaponization
  - Emphasizes need for analysis of unanticipated emergent behavior.
- **(Oct 2013) DoD, *Unmanned Systems Integrated Roadmap: FY2013-2038***[151]
  - Outlines three broad goals: (1) expand scope of unmanned systems and integrate them into the current military structure, (2) be cost-effective, as defense budgets are cut (e.g., drone programs were cut 33% between

[149] http://www.nrac.navy.mil/docs/NRAC_Final_Report-Autonomy_NOV2012.pdf.

[150] http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

[151] http://archive.defense.gov/pubs/DOD-USRM-2013.pdf. The next version of DoD's unmanned systems roadmap is scheduled to be completed the first quarter of fiscal year 2017, and will examine systems in the ground, air and maritime domains for FY-16 through FY-41. Ref: J. Mishory, "DOD preparing to release new 25-year unmanned systems roadmap," *Inside Defense*, 27 Oct 2016

2013 and 2014), and (3) build drones that can operate in less permissible environments than Iraq and Afghanistan.

  - Has 65/120 pages of text devoted to discussions of technology for unmanned systems.
  - Identifies *interoperability* as one of the principal needs in improving the use of unmanned systems.

- **(Nov 20143) SecDef (Chuck Hagel) announces new *Defense Innovation Initiative: Third "Offset" Strategy* at the Reagan National Defense Forum Keynote**[152]
  - Announces new Defense Innovation Initiative: Third "Offset" Strategy.
  - Discusses five key points of strategy: (1) autonomous "deep learning" systems, (2) human-machine collaboration, (3) assisted-human operations, (4) human-machine teaming, and (5) semi-autonomous weapons."
- **(June 2015) Memorandum, SecNav*: AI and Robotics for Support Functions***[153]
  - Cites recent operational examples of the application of AI and robotics technology within the Navy—e.g., *Ghost Swimmer*, an underwater unmanned vehicle that mimics a bluefin tuna,[154] the X-47B, that can autonomously land aboard an aircraft carrier,[155] and the *Swarmboat* USV, that can synchronize with other unmanned vessels to swarm.[156]
  - Recognizes heavy private sector investment in AI and robotics.
  - Mandates acceleration of exploring of these emerging fields; i.e., for DoN to identify opportunities for the DON integration of proven AI and robotics technologies.
- **(June 2015) *Human Autonomy Teaming*, United States Air Force, Office of the Chief Scientist**[157]

[152] http://www.defense.gov/News/Speeches/Speech-View/Article/606635.

[153] http://www.secnav.navy.mil/innovation/Documents/2015/06/AIRoboticsMemo.pdf.

[154] J. Golson, "The Navy's new robot looks and swims just like a shark," *Wired*, 16 Dec, 2014.

[155] X-47B UCAS Demo Evolution: http://www.navair.navy.mil/img/uploads/UCASTimeline%20(2).png.

[156] S. Smalley, "The Future Is Now: Navy's Autonomous Swarmboats Can Overwhelm Adversaries," Office of Naval Research: https://www.onr.navy.mil/Media-Center/Press-Releases/2014/autonomous-swarm-boat-unmanned-caracas.aspx.

[157] *Autonomous Horizons: System Autonomy in the Air Force A Path to the Future, Volume I*: Human-Autonomy Teaming, United States Air Force, Office of the Chief Scientist, June 2015. Volumes II (that will deal with technical issues in creating intelligent machines that can operate in uncertain and changing environments) and III (that will address cyber security and reliability,

  - Articulates Air Force's vision for autonomous systems that "serve as a part of a collaborative team with airmen. Flexible autonomy will allow the control of tasks, functions, sub-systems, and even entire vehicles to pass back and forth over time between the airman and the autonomous system, as needed to succeed under changing circumstances. Many functions will be supported at varying levels of autonomy, from fully manual, to recommendations for decision aiding, to human-on-the-loop supervisory control of an autonomous system, to one that operates fully autonomously with no human intervention at all."
  - Identifies key technical challenges for human-machine teaming: (1) developing autonomous systems that are robust enough to function without human intervention and oversight; (2) spectre of reducing situation awareness (by inadvertently leaving humans out-of-the-loop); (3) potential increase in cognitive load of humans (while interfacing with autonomous systems); (4) decision speed vs. decision accuracy tradeoffs; (4) establishing trust (that is appropriately calibrated to the reliability and functionality in given operational contexts).

- **"(Feb 2016) *Human Systems Roadmap Review*, Office of SecDef, Human Systems Community of Interest (HSCOI)**[158]
  - Highlights need for: "More robust, valid, and integrated mechanisms that enable constructive agents that truly think and act like people."
  - One of the five "building blocks" of the Human Systems program is to develop: "Network-enable, autonomous weapons hardened to operate in a future Cyber/EW Environment" … [and allow for] "…cooperative weapon concepts in communications-denied environments."
  - Focus areas for S&T development include: (1) "autonomous weapons: systems that can take action, when needed," and (2) "architectures for autonomous agents and synthetic teammates."

- **(Feb 2016) *Autonomous Undersea Vehicle (AUV) Requirements for 2025*, Report To Congress, Chief of Naval Operations, Undersea Warfare Directorate**[159]

communication links, and command and control systems to support autonomous vehicles) have not yet been released (as of Dec 2016).

[158] HSCOI consists of senior officials from the US Army, Navy, Marine Corps, Air Force, Defense Advanced Research Projects Agency (DARPA); and is overseen by the Assistant Secretary of Defense for Research & Engineering and the Assistant Secretary of Defense for Health Affairs; http://www.defenseinnovationmarketplace.mil/resources/NDIA_Human_Systems_Conference_2016_HSCOI_DistroA_FINAL.pdf.

[159] https://news.usni.org/wp-content/uploads/2016/03/18Feb16-Report-to-Congress-Autonomous-Undersea-Vehicle-Requirement-for-2025.pdf.

  - Identifies AUVs as key component to expand undersea superiority.
  - Projects that by 2025 AUVs will operate farther forward than manned platforms; will operate in shallower, denied waters; will support multiple tasks/sensors; will be more passive vice transmitting frequently; and will have increased general autonomy (e.g., managing own stealth)
  - Delineates specific tasks that (by 2025) AUVs will be able to perform independently or cooperatively (enhancing) tasks performed by manned submarines.

- **(June 2016) Defense Science Board (DSB) Task Force Report (TF): *Summer Study on Autonomy***[160]
  - Autonomy has "attained a "tipping point' in value": "DoD must take immediate action to accelerate its exploitation of autonomy while also preparing to counter autonomy employed by adversaries."
  - Multiple recommendations aligned with three over-arching vectors: (1) accelerating DoD's adoption of autonomous capabilities, (2) strengthening the operational pull for autonomy, and (3) expanding envelope of technologies available for DoD missions.
  - Recommendations are grounded on in concepts of "autonomy" and "trust."

- **(Oct 2016) *The National Artificial Intelligence Research and Development Strategic Plan*, National Science and Technology Council**[161]
  - Identifies AI as a transformative technology.
  - Defines objectives for federally-funded AI research: (1) acquiring long-term investments in AI research; (2) developing effective method for human-AI collaboration; (3) understanding ethical and legal implications
    of AI; (4) ensuring safety and security of AI systems; (5) developing shared public datasets for AI training and testing; (6) establishing standards and benchmarks for evaluating AI technologies; and (7) better understanding national AI research and development needs.

[160] http://www.acq.osd.mil/dsb/reports/DSBSS15.pdf.

[161] http://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf.

What do they all have in common?

Figure 9 summarizes the main technologies mentioned in each of the reports listed above, and identifies the eight most commonly cited ones in the center.

Figure 9. Core set of innovations/technologies across a representative set of recent studies and reports (references appear on the next page)

[1] *Emerging Science Technologies*, Intelligence and National Security Alliance, Council on Technology and Innovation, April 2013.

[2] S. Brimley, K. Sayler, and B. Fitzgerald and, "Game Changers: Disruptive Technology and U.S. Defense Strategy", Center for a New American Security (CNAS), Sep 2013.

[3] *Envisioning 2030: US Strategy for the Coming Technology Revolution*, Strategic Foresight Initiative, Brent Scowcroft Center on International Security, 2013.

[4] R. Work and S., Brimley, *20YY: Preparing for War in the Robotic Age*, CNAS, Jan 2014.

[5] *Summer Study on Autonomy*, DoD, Defense Science Board (DSB), Task Force Report (TFR), Office of the Under SecDef for Acquisition, Technology and Logistics, June 2016.

[6] *The Role of Autonomy in DoD Systems*, DoD, DSB, TFR, Office of the Under SecDef for Acquisition, Technology and Logistics, July 2012.

[7] *The National Artificial Intelligence Research and Development Strategic Plan*, National Science and Technology Council, Networking and Information Technology R&D Subcommittee, Oct 2016.

[8] *Robotics Collaborative Technology Alliance: FY 2012 Annual Program Plan*, Army Research Laboratory, March 2012.

[9] *Unmanned Systems Integrated Roadmap: FY2013-2038*, U.S. Department of Defense: http://archive.defense.gov/pubs/DOD-USRM-2013.pdf.

[10] *Global Trends 2030*, National Intelligence Council, NIC 2012-001, Dec 2012.

[11] *Technology and Innovation Enablers for Superiority in 2030*, DoD, DSB, TFR, Office of the Under SecDef for Acquisition, Technology and Logistics, Oct 2013.

[12] http://www.defenseinnovationmarketplace.mil/resources/Lemnios_Testimony_2013.pdf.

Artificial intelligence

What is it?

Despite its long history, and increasingly rapidly paced advances, there has never been a universally agreed-upon definition of what "AI" is. Norvig and Russell, in the introduction to their opus *Artificial Intelligence: A Modern Approach*,[162] provide extracts of eight different definitions of AI from other standard textbooks, some of which stress "thought processes," others "reasoning," and still others "rationality"; (The authors point out that proponents of these various approaches have both helped and disparaged each other.) Nilsson, in the preface of his recent historical survey of the field—*Quest for Artificial Intelligence*[163]—defines AI as "that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment." Poole and Mackworth provide an even simpler definition, albeit one laden with suggestive concepts:[164] "AI is the field that studies the synthesis and analysis of computational agents that act intelligently."

The "Turing Test," proposed by Alan Turing in 1950,[165] was designed to provide an operational means of detecting the presence of intelligence in an engineered system. Turing's idea was to use an existing "standard" of intelligence (i.e., that of a human) to probe for the ability of an AI-candidate achieve the same level of performance in the same set of cognitive tasks that the "standard" is proficient in. The method was to have a human interrogate the candidate system (in Turing's time, via a teletype). If the interrogator cannot determine whether she is probing a computer or a human, the candidate "passes" the test. Leaving aside the question of whether the "AI" that

[162] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, 2009.

[163] N. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge University Press, 2009.

[164] D. Poole and A. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*, Cambridge University Press, 2010.

[165] A. Turing, "Computing Machinery and Intelligence," *Mind* 236, Oct 1950.

passes such a test is intelligent,[166] even the minimal set of capabilities that a system would need in order to pass such a test effectively covers most of what today constitutes the field of AI:

- *Natural language processing*, so that the system can communicate with the interrogator (modern incarnations are coupled with automatic speech recognition).[167]
- *Knowledge representation*, to codify what the system knows.
- *Automated reasoning*, so that the system has the facility to use its stored information to draw inferences and answer questions.
- *Machine learning*, so that the system is able to adapt to new information, by incorporating news facts into its existing corpus and to detect (and extrapolate from) patterns of information.

While Turing deliberately left out a requirement for the interrogator to physically interact with the candidate system (since "intelligence" does not require corporeality),[168] one can imagine a generalized "Total Turing test" that includes, say, a video feed to allow the interrogator to test a candidate AI's perceptual skills. In this case, the candidate system would need additional capabilities:

- *Computer vision*, to sense objects in its immediate environment (and the currently dominant form of machine perception); today, for some specific image classification tasks, computers perform *better* than humans.[169]
- *Robotics*, in order to manipulate and otherwise interact with objects. Current research focuses on interacting with changing environment (the general navigation problem having been effectively "solved" for static environments).[170]

[166] See chapter 26 in S. Russell and P. Norvig, *Artificial Intelligence.*

[167] Near-real-time translation already commercially available: M. Guta, "Waverly Labs Earpieces Translate Other Languages Almost Instantly," *Small Business Trends*, 30 May 2016.

[168] The fact that modern cognitive theory posits that intelligence (at least as embodied within humans) is actually entwined with physicality, does not alter our narrative, since Turing was concerned only with whether a *concealed* AI-system (i.e., an algorithm running on a computer that is itself situated somewhere else; e.g., "behind a curtain") can fool a human into believing she is interrogating another human. Ref: C. Allen, "Why Intelligence Requires Both Body And Brain," *Footnote*, 27 Jan 2014.

[169] A. Hern, "Computers now better than humans at recognizing and sorting images," *The Guardian*, 13 May 2015.

[170] R. Murphy, *An Introduction to AI Robotics*, MIT Press, 2000.

State-of-the-art advances in many of these fields (particularly computer vision and natural language processing) have been spurred by recent advances in deep learning techniques (discussed later). Other forms of AI (not conceivable in Turing's time), include *collaborative systems* (i.e., autonomous systems that work alongside humans and other AI systems), *crowdsourcing* (i.e., harnessing human intelligence, such as *Wikipedia* does), and *Internet-of-Things* (i.e., the interconnected world of physical devices that can collect and share their sensory information). To make these otherwise abstract concepts more concrete and relevant in the context of the subject matter of this report, the next few sections take a deeper dive.

Overview

Research and development in AI officially dates back to the 1956 British Dartmouth Summer Conference at which John McCarthy coined the term,[171] although its rudiments—at least as far as inspiration from biological neural processing is concerned—go back at least a decade to the introduction of the first artificial neuron (or threshold logic unit), introduced in 1943 by McCulloch and Pitts.[172] Since then the field has advanced along two concurrent methodological points of view: a *top-down* approach, in which knowledge about a specific problem domain is first curated by human subject matter experts (SMEs), codified in terms of simple rules, and implemented in software (the goal of which is to reproduce "human like" reasoning); and a *bottom-up* approach, which deliberately mimics nature's own evolutionary propensity to build "complex" structures out of "simpler" parts. Whereas the first approach generally seeks to create AI systems that "understand" (a segment of) the world by imposing a hand-crafted symbolic ontology (i.e., a semantic model that describes a SME-curated knowledge), the latter approach is grounded on the belief that AI systems must learn to understand their environments (and problem domains) on their own. The best known examples of these two approaches are, respectively, *expert systems* (ESs) and *machine learning* (ML). A third approach, *natural language processing* (NLP), involves aspects of both ES and ML and is playing an increasingly central role in advancing the state-of-the-art in human-AI collaboration.[173]

[171] John McCarthy is generally acknowledged as one the "founding fathers" of artificial intelligence, along with Marvin Minsky (with whom he worked at MIT), Allen Newell, and Herbert Simon; Ref: S. Williams, *Arguing A.I.: The Battle for Twenty-first-Century Science*, AtRandom Books, 2002.

[172] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, Vol. 5, 1943.

[173] R. Deits, et al., "Clarifying Commands with Information-Theoretic Human-Robot Dialog," *Journal of Human-Robot Interaction* 2, 2013.

Research in NLP goes back to the roots of computer science in the 1940s and 1950s.[174] Today it is a mix of computer science, general AI and computational linguistics, with a specific focus on the automatic "understanding" of free-form human language. NLP, by itself, does not denote any specific method or algorithm, but is best thought of as a label for a broad rubric of related techniques. Examples include: *text summarization*, in which a given document is distilled to manageably small summary; *named entity recognition* (NER), which is the task of identifying text elements that belong to certain predefined categories, such as the names of persons, organizations, locations, and expressions of times; *relationship extraction*, in which the relationship between various named parts of a chunk of text are identified ("an object O *belongs* to person P"); *semantic disambiguation*, in which a priori ambiguous meanings of words (or chunks of text) are automatically disambiguated from a deeper analysis of context and/or information that may be culled from an "ontology" (see discussion below); *sentiment analysis*, in which certain kinds of subjective information is extracted from a document or set of documents (e.g., extracting a range of emotional reactions to public events from social media posts); *speech recognition*, which refers to the textual representation of sound recordings of people speaking; and *natural language understanding*, in which semantic content is extracted from free-form text (this is arguably "the" most difficult open-research problem of NLP).

NLP consists of myriad subfields, including: (1) *machine translation* (the automatic translation from one language to another); (2) *information retrieval* (the act of obtaining, storing, searching information resources that are relevant to a specific query or subject from a given source of documents); (3) *information extraction* (the extract of semantic information from text); and (4) *deep learning* (discussed below).

Expert systems

ESs may be characterized loosely as systems that mimic the decision-making ability of a human expert (in a given domain), and are arguably the best known "successes" of early 1970s AI research.[175] They are essentially a knowledge base (see "Semantic Knowledge Models" below) + inference engine, with the latter typically instantiated as a set of SME curated "IF…THEN…" rules. For example, DENDRAL[176] (developed by

[174] S. Lucci and D. Kopec, *Artificial Intelligence in the 21st Century*, Mercury Learning and Information, 2013.

[175] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, 2009.

[176] E. Feigenbaum, B. Buchanan, and J. Lederberg, "On generality and problem solving: a case study using the DENDRAL program," *Machine Intelligence* 6, 1971.

Stanford University, 1965-1969) was designed to deduce the molecular structure from the information provided by a mass spectrometer; and MYCIN[177] (also developed at Stanford University in the early 1970s) was an automated system to diagnose infectious diseases. It was among the first expert systems to combine SME "rules" (about 450) with a calculus that allowed the system to make inferences based on "certainty factors," and performed better than junior doctors. The first commercial expert system, R1[178] (introduced by the Digital Equipment Corporation in 1982), was designed to help managers configure orders for their new VAX-11 computer systems, and was powerful enough to save the company about $40 million a year). Of course, numerous other examples may be culled from the literature, and the range of applications is both deep and broad. For example, a 1983 book by Hayes-Roth, Waterman, and Lenat[179] lists 10 general categories of applicability of ES technology: (1) interpretation and identification (i.e., making general inferences from input data), (2) prediction, (3) diagnosis, (4) design, (5) planning, (6) monitoring, (7) debugging, (8) repair (e.g., developing a plan to administer a required remedy to a system fault), (9) instruction, and (10) control.

MYCIN, because of its early success, remains a benchmark against which even modern ES are often judged. Its importance is underscored by Durkin,[180] among whose "lessons learned" from MYCIN are: *knowledge is separate from control* (i.e., one does not need to change the inference engine if the rules change); *rules must accommodate both inexact reasoning* (i.e., degrees of certainty) *and meta-control* (i.e., rules about rules); *natural language interaction* (i.e., the user must be able to interact with the system in a natural fashion); *explanatory power* (i.e., the system must be able to explain how and why a given inference was made); and an ability to provide alternative recommendations (so that the user is not constrained to blindly choose a single "conclusion"). An equally important lesson is that despite being a relatively "simple" expert system by modern standards, MYCIN was hardly "easy" to develop, requiring some 20-person years of research effort. The same is true of modern approaches to the "general AI" problem, examples of which will be discussed below.

[177] B. Buchanan and E. Shorti, *Rule based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, 1984.

[178] J. Macdermott, "R1: the formative years," *AI Magazine* 2, no. 2, 1981.

[179] F. Hayes-Roth, D. Waterman, and D. Lenat, *Building Expert Systems*, Addison-Wesley, 1983.

[180] J. Durkin, *Expert Systems: Design and Development*, Macmillan, 1994.

Machine learning

"Machine learning" is a catch-all phrase that refers to a wide variety of techniques designed *to detect patterns in, and learn and make predictions from data*. Specific techniques include:[181] *Bayesian belief networks* (which are graph models whose nodes represent some objects or states of a system and whose links denote probabilistic relationships among those nodes); *deep learning* (which is sometimes also called *hierarchical learning*, and refers to a class of ML algorithms designed to find multiple high levels of abstract representations of patterns in data); *genetic algorithms* (and other *evolutionary programming* techniques that mimic the dynamics of natural selection);[182] *inductive logic programming* (designed to infer a hypothesis from a knowledge base and a set of positive and negative examples);[183] *neural networks* (which are inspired by the structure and function of biological neural networks);[184] *reinforcement learning* (which is inspired by behaviorist psychology and refers to a technique whereby learning proceeds by adaptively constructing a sequence of actions that collectively maximize some long-term reward);[185] and *support vector machines* (SVM),[186] used for classifying objects into While all ML techniques require a dataset (or multiple datasets) to be used as a source of training data, the learning can proceed in one of three ways: *supervised*, *semi-supervised*, or *unsupervised*. In *supervised learning*, each training data element is explicitly labeled as an input-output pair, where the output is the "correct" desired value that one wishes the system to learn to associate with a given input (thereby learning the general rules by which to associate input-output pairs not in the original training set), and the "output" represents a "supervisory signal"). In *unsupervised learning*, the system attempts to discover hidden structure in data on its own—i.e., no reward signals are given to "nudge" the system as it processes the training data. *Semi-supervised learning* refers to a class of supervised learning techniques that also use unlabeled training data. Reinforcement learning may be considered a form of semi-supervised learning, in that it neither uses input-output pairs for training nor is completely unsupervised; instead, the type of feedback it receives depends on its

[181] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, 2009.

[182] Z. Michalewicz and D. Fogel, *How to Solve It: Modern Heuristics*, Springer-Verlag, 2005.

[183] S. Muggleton and H. Watanabe, *Latest Advances in Inductive Logic Programming*, Imperial College Press, 2014.

[184] M. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, 2003.

[185] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[186] N. Cristianini, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

response. For correct responses, it receives the same type of response as any supervised learning system does (e.g., response is "correct"); for incorrect responses it is told only that an "incorrect response" was given but is not informed of what the correct response was.

Neural networks and deep learning

Neural networks (NNs) are among the oldest, and most powerful, forms of "bottom up" AI methods.[187] Though the development of the general method was curtailed during at least two dark periods (in the 1970s and 1990s; see below), NNs are currently undergoing a burgeoning renaissance in a slightly modified form known as *Deep Learning* (DL),[188] due mainly to the confluence of three factors: exponential growth in computing power, the exponentially dwindling cost of digital storage coupled with an exponential growth of available data, and a new generation of fast learning algorithms for multiplayer networks. Because of the importance of these techniques to the development of military autonomous systems and AI in general, one must have at least a passing acquaintance with the history of NNs and basic terminology in order to appreciate the significance of the most recent developments. Figure 10 shows a timeline of major milestones, starting with the first mathematical model of a neuron introduced in 1943.[189]

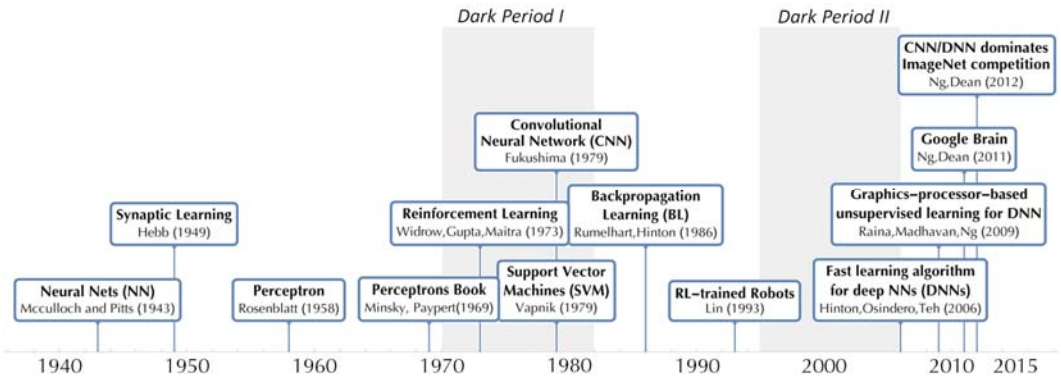
At its core, and loosely stated, an NN represents a particular class of functional transformations from a set of input patterns to an output class of associated categories. Think of a simple linear regression (LR)—say, a linear predictor function (LPF)—for modeling the relationship between some scalar dependent variable, y, and a single independent variable, x. To find the LPF, one merely has to fit a line that "best fits" the dataset that represents what is known about how y is related to x (e.g., the dataset may consist of a set of (x,y) pairs, most, or even all, of which may only be known approximately). Of course, one is free to use a more complicated function, but this runs the risk of *overfitting* (i.e., "learning" a function that works well for the data in the training set but is unable to predict reasonable sets of values for the "real" or "test" data).

[187] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Technical Report IDSIA-03-14*, arXiv:1404.7828 v4, 2014.

[188] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[189] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics* 5, 1943.

Figure 10. Timeline of milestones in the development of neural networks and deep learning techniques (see text for discussion)



LR is essentially the idea behind the *perceptron*, introduced by Rosenblatt in 1958 as a mathematical distillation of how biological neurons operate.[190] In a perceptron, each "neuron" takes a set of binary inputs (from nearby neurons in the NN), multiplies each input by some real-valued weight (which represents the strength of the connection to each nearby neuron), transforms the sum of these weighted inputs to an output value of 1 if the sum exceeds some threshold value, and otherwise outputs the value 0 (to mimic the way biological neurons either "fire" or not). It was believed, early on, that perceptrons could be used as the basis for developing AI systems because it can be proven that they can model basic logic functions (such as OR, AND, and NOT gates). In order to "learn" a function, one starts with an input-output training set, and adjusts the weights of the perceptron by either increasing their value if the output for a given example is too low, or decreasing their value if the output is too high. The rudiments of modern ML were born when Rosenblatt's learning-perceptron was implemented in hardware (and used to classify simple shapes with 20-by-20 pixel inputs), and the single-output perceptron design was replaced with a network that included multiple neurons in the output layer. For example, in the latter case, if the task is for the NN to "learn" to classify an image of a handwritten digit, the inputs may be used to represent the pixels of an image, and 10 output neurons may be used to correspond to each of the 10 possible digit values.

The first "dark period" of NN development (see figure 10), during which the funding for further research and the number of published papers dropped significantly from

[190] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, Vol. 65, 1958.

prior years, followed the landmark publication of the book *Perceptrons* in 1969.[191] The book argued, correctly, that because the Boolean exclusive-OR (or XOR) function is not linearly separable,[192] the utility of *Perceptron* networks in the development of AI is necessarily limited. The only way that the XOR function can be learned was by a multilayer network; i.e., a NN in which there are *hidden layers* sandwiched between the input and output layers (see left-hand-side of figure 11-a).

But the only known learning algorithm at the time of the book's publication applied only to the simplest single-input-layer/single-out-layer NNs. It was not until 1986, when the so-called *Backpropagation Learning* (BL) method, which could be applied to NNs with hidden layers, was introduced, that the first "dark period" of NN research finally ended.[193] Three years later, Hornik et al. proved that multiply layered NNs can learn any function, including XOR.[194] Also in 1989 was the first landmark application of BL to the automatic recognition of handwritten zip code numbers;[195] the algorithm for which was a precursor of what have come to be known as *convolutional neural networks* (CNNs). The layers of a CNN are defined to exploit any regularities and constraints of the dataset that the NN is being trained on. For example, if the NN is to be trained on a set of 3D images, the layers of a CNN might be arranged in three dimensions (width, height, and depth).[196] Modern incarnations of CNNs include *pooling layers* (positioned in-between convolutional layers), that effectively reduce the spatial representation (e.g., by down sampling the size of an image) to reduce the number of parameters in the network, and thus also help control overfitting.

[191] M. Minsky and S. Paypert, *Perceptrons*, MIT Press, 1969.

[192] The XOR function yields the following output values for the four possible input combinations of 0 and 1: 0 XOR 0 = 0, 0 XOR 1 = 1, 1 XOR 0 = 1, and 1 XOR 1 = 0. If these four output values are arranged in a two-dimensional (x,y) plot, it is immediately clear—by visual inspection—that it is impossible to draw a line that separates the two '0' values from the two '1' values. Ref: Chapter 4.5 in S. Haylin, *Neural Networks: A Comprehensive Foundation*, Second Edition, Prentice Hall, Inc., 1999.

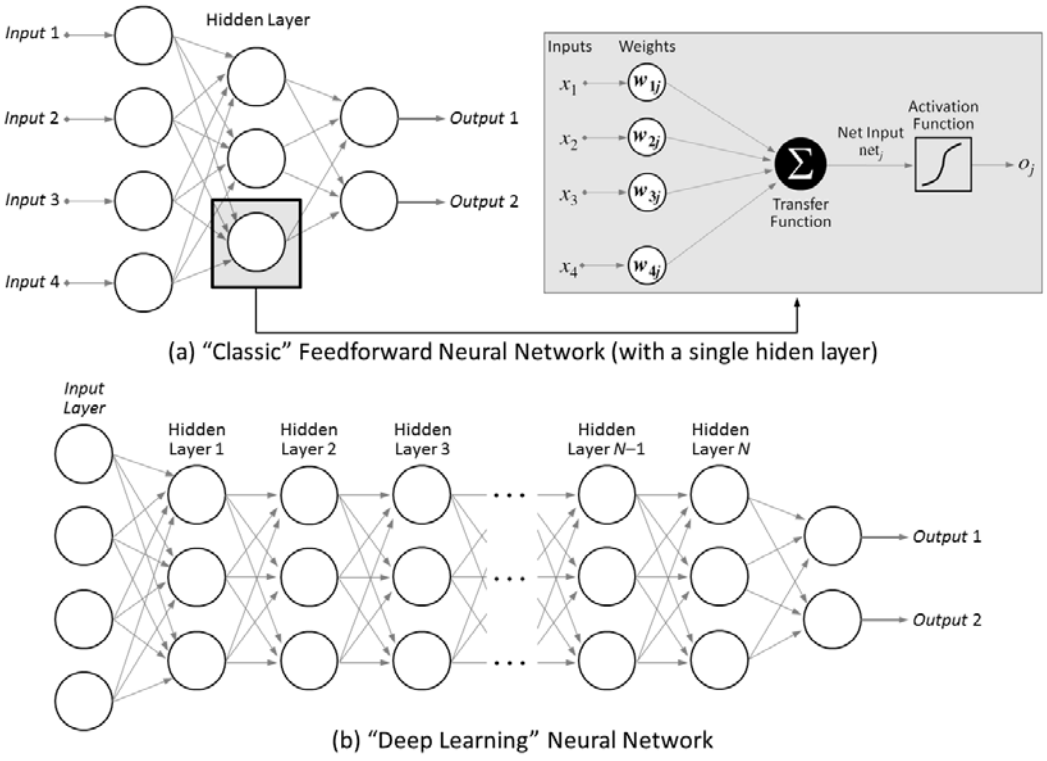
[193] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature* 323, 1986.

[194] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks* 2, 1989.

[195] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation* 1, 1989.

[196] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

Figure 11. Schematic illustrations of neural network designs



The late 1980s/early 1990s saw the advent of NNs being applied to physical robots, in which, for example a robot was taught (using supervised learning) to steer through a simple physical environment.[197] Also around this time, an RL-based AI system—called TD-Gammon—famously "taught itself" to play backgammon at a superhuman level.[198] TD-Gammon is one of the first instances of an RL/NN-hybrid system being able to outperform humans on a relatively complex task (see discussion in next section). Ironically, it was this early "success" that led to the second "dark period" of NN development (see figure 10), which ended relatively recently in 2006. The reason was that when TD-Gammon's learning algorithm was applied to other

[197] L. Lin, *Reinforcement learning for robots using neural networks*, Ph.D. Thesis, Carnegie-Mellon University, School of Computer Science, CMU-CS-93-103, 1993.

[198] G. Tesauro, "Temporal difference learning and TD-Gammon," *Communications of the ACM* 38, 1995.

(albeit more "complex") games such as chess[199] and Go,[200] its performance was far worse. Notably, the main reason for the ostensible "failure" was not so much the learning algorithm (the basic characteristics of which are still embedded in most modern "successes"; see next section), but the relative *slowness* of the computer processors and *limited memory storage* of c.1990s era computers. Amidst the growing realization that problems more "complex" than that of learning to play backgammon required many more than one single hidden layer, was the reality that the BP algorithm did not work well for NN that had many hidden layers[201]—and it is the presence of many hidden layers that is the cornerstone of most modern deep learning systems (see figure 11-b).[202]

Even as the raw power of available computers was steadily increasing until very recently (thanks to Moore's law[203]), it was not until both a "fast learning" algorithm for deep learning neural networks (DLNNs) was finally introduced in 2006[204] and AI researchers began exploiting the massively parallel computing powers of Graphical Processing Units (GPUs) to speed up learning even further,[205] that a large—and increasing—number of "narrow AI" problems showed signs of having been effectively "solved" (see **State-of-the-Art** below).

---

[199] S. Thrun, "Learning to play the game of chess," *Advances In Neural Information Processing Systems* 7, 1995.

[200] N. Schraudolph, P. Dayan, and J. Sejnowski, "Temporal difference learning of position evaluation in the game of Go," *Adv. in Neural Information Processing Systems* 6, 1994.

[201] A. Ilachinski, Chapter 10 in *Cellular Automata*, World Scientific Press, 2001.

[202] The BP (supervised) learning rule is essentially a prescription for adjusting the initially randomized set of synaptic weights (existing between all pairs of neurons in each successive layer) so as to minimize the difference between the perceptron's output of each input fact and the output with which the given input is known (or desired) to be associated. The backpropagation rule takes its name from the way in which the calculated error at the output layer is propagated backwards from the output layer to the N^{th} hidden layer to the $(N-1)^{th}$ hidden layer, and so on. As the number of layers, N, increases, the BP rule results in assigning unmanageably large or extremely small numbers to weights; i.e., the 'vanishing or exploding gradient problem." Ref: J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks* 61, 2015.

[203] "Moore's law" was an observartion made in 1965 by Gordon Moore, co-founder of Intel, that the overall processing power for computers roughly doubles every two years; a pattern that has only recently been broken: T. Simonite, "Moore's Law Is Dead. Now What?", *MIT Technology Review*, 13 May 2016.

[204] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation* 18, 2006.

[205] Speedups close to *two orders of magnitude* have been reported: R. Raina, A. Madhavan, and A. Ng, "Large-scale deep unsupervised learning using graphics processors," *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009.

While the United States government has always played an important role in fostering AI research (e.g., ARPA, DARPA, NSF, ONR),[206] it is arguably the commercial sector that is driving most significant advances (certainly for "narrow AI" problems), including major technology companies such as Amazon, Apple, Microsoft, Google, and Facebook.[207] Nearly 140 private AI technology companies have been acquired since 2011, with over 40 acquisitions taking place in 2016 alone (as of this writing: 27 October 2016).[208] Perhaps the most noteworthy acquisition (anticipating the discussion of some major AI "successes" in the next section) occurred in 2014, when Google bought the AI startup *DeepMind*. More recently, both Microsoft and Google have recently made public essentially the same set of core machine learning tools used by their in-house researchers: *Computational Network Toolkit* (CNTK)[209] and *TensorFlow*,[210] respectively.

Finally, we note the United States is no longer the "de facto" world leader when it comes to the number of publications' citations of research journals mentioning "deep learning" or "deep neural network"; that distinction, as of sometime between 2013 and 2014, now belongs to mainland China (see figure 12).[211]

[206] *Funding a Revolution: Government Support for Computing Research, National Research Council*, National Academy Press, 1999.

[207] "Microsoft, Google, Facebook and more are investing in artificial intelligence: What is their plan and who are the other key players?", *TechWorld*, September 29, 2106, http://www.techworld.com/picture-gallery/big-data/9-tech-giants-investing-in-artificial-intelligence-3629737/.

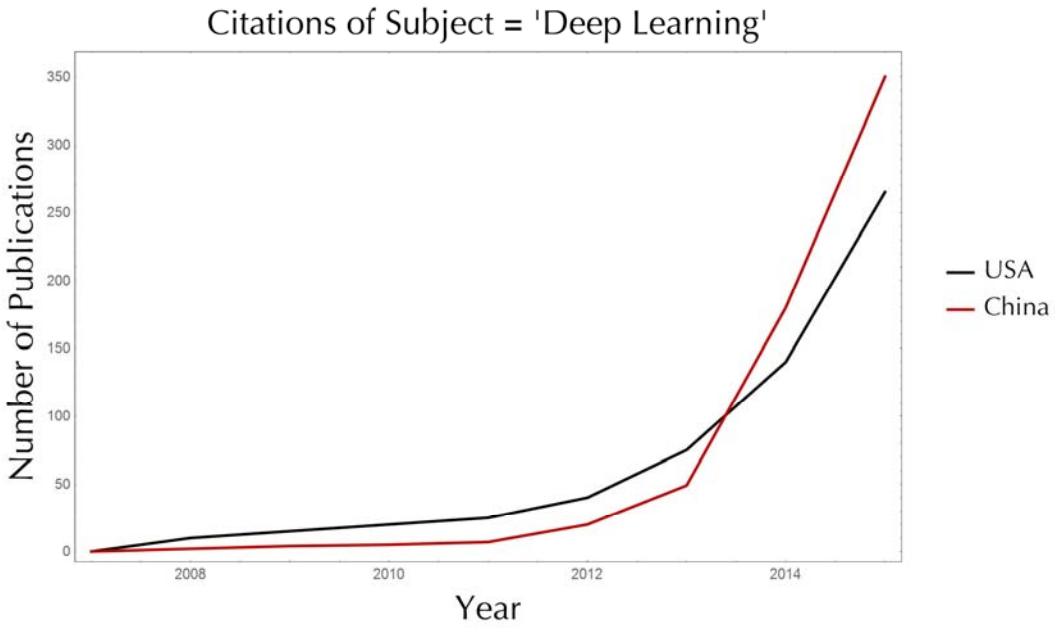
[208] The Race For AI: Google, Twitter, Intel, Apple In A Rush To Grab Artificial Intelligence Startups, *CBInsights*, 7 Oct 2016: https://www.cbinsights.com/blog/top-acquirers-ai-startups-ma-timeline/.

[209] https://github.com/Microsoft/CNTK.

[210] https://github.com/tensorflow/tensorflow.

[211] *The National Artificial Intelligence Research and Development Strategic Plan*, National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, Oct 2016.

Figure 12. Number of journal articles mentioning "deep learning" or "deep neural network" for the top 6 nations (as of 2015)



Ref: Based on Figure 1 in Ref: The National Artificial Intelligence Research and Development Strategic Plan, National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, Oct 2016.

State-of-the-Art

Increasingly more often, AI systems are outperforming humans on *specific tasks*; a fact due primarily to the advent of deep learning techniques (discussed above) and the accompanying (and continuing) growth of computer power. Table 2 shows selected milestones when AI first surpassed human performance, culminating in Google's DeepMind's *AlphaGo* defeating a top-ranked human Go player (Lee SeDol) four games to one in March 2016.[212]

[212] http://gogameguru.com/alphago-defeats-lee-sedol-4-1/. *AlphaGo* also defeated Fan Hui, the European Go champion, five games to none in October 2015. However, the 2016 match against Lee Sedol was the first time that a human with the highest Go ranking (9-dan master) lost a tournament match without handicap to an AI system.

Table 2. Selected milestones when AI first surpassed human performance

Year	Milestone	Reference
1981	*Traveller TCS**	D. Lenat, "EURISKO: A program that learns new heuristics and domain concepts," *Artificial Intelligence* 21, 1983
1992	*Backgammon*	G. Tesauro, *A Self-Teaching Backgammon Program, Achieves Master-Level Play*, AAAI Technical Report FS-93-02, 1993
1994	*Checkers*	J. Schaeffer, *One Jump Ahead: Challenging Human Supremacy in Checkers*, Springer-Verlag, 1997
1997	*Othello*	M. Buro, "The Othello match of the year: Takeshi Murakami vs. Logistello," *ICGA Journal* 20, 1997
	Chess	B. Pandolfini, "Kasparov and Deep Blue: The Historic Chess Match Between Man and Machine," *Fireside Chess Library*, 1997
2002	*Scrabble*	B. Sheppard, "World Championship Caliber Scrabble," *Artificial Intelligence* 143, 2002
2008	*Poker*	J. Rubin, I. Watson, "Computer poker: A review," *AI* 175, 2011
2011	*Trivia/Jeopardy!*	S. Baker, *Final Jeopardy: Man vs. Machine and the Quest to Know Everything*, Houghton Mifflin Harcourt, 2011
2013	*Atari Games*	V. Mnih et al., "Asynchronous Methods for Deep Reinforcement Learning," in *Proc. of the 33rd International Conference on Machine Learning*, 2016
	Image Recognition	K. He et al., "Deep Residual Learning for Image Recognition," *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016
	Speech Recognition	D. Amodei, "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin," *Proceedings of the 33rd International Conference on Machine Learning*, 2016
2016	*Go*	S. Byford, "Google's AlphaGo AI beats Lee Se-Dol again to win Go series 4-1," *The Verge*, March 15, 2016
	Voice Recognition	W. Xiong et al., "Achieving Human Parity in Conversational Speech Recognition," 17 Oct 2016 (http://arxiv.org/abs/1610.05256v1)

* "Traveller TCS" is a futuristic naval war game, introduced in 1977 by Game Designers' Workshop.

Although IBM's *Deep Blue*'s victory in chess over Gary Kasparov in 1997 was historic at the time, *AlphaGo*'s victory in Go over Lee Se-Dol in 2016 is arguably more noteworthy, because of the generalizability of the learning method by which it was taught). Indeed, right up until *AlphaGo*'s victory, it was commonly believed among AI researchers that it would take at least another decade for an AI to defeat a top-ranked Go player.[213] The reason simply has to do with how much more "complex" a game Go is compared to chess. For example, using one standard metric, called *game tree complexity* (GTC)—which, roughly speaking, measures the number of positions a move-ranking algorithm would have to evaluate in order to determine the value of an initial position—the complexity of Go exceeds that of chess by almost 240 *orders of magnitude*.[214]

[213] http://www.wired.com/2016/01/in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/.

[214] While chess is played on an 8-by-8 board, tournament-level Go is played on a 19-by-19 board, albeit with effectively two pieces. J. Burmeister, "The challenge of Go as a domain for AI research: a comparison between Go and chess," *Intelligent Information Systems*, 1995.

For purposes of this white paper, it is instructive to underscore the *differences* between *Deep Blue*'s and *AlphaGo*'s respective achievements. While both AI systems obviously learned to play (chess and Go, respectively), the ways in which they were "taught"—and how they *played*—are very different. *Deep Blue*'s core evaluation function numerically "ranks" given board positions, and was handcrafted, albeit with many thousands of open parameter values, and later refined by a grandmaster. The final algorithm was effectively determined by the system itself, after playing several thousand games against itself. The style of gameplay (which defeated Kasparaov) was effectively "brute-force," in which *Deep Blue* systematically applied its evaluation function to many alternative future states, searching seven or eight moves ahead for each player, at a rate of about 200 million position evaluations per second.[215] Since *Deep Blue*'s victory over Kasparov in 1997, chess-playing computers have become increasingly stronger, to the point where all but the strongest players are likely to be defeated by chess engines running on a smartphone.[216] However, the manner in which these chess engines "play" has not changed, as they all rely on the same "brute force" approach used by *Deep Blue*.

AlphaGo's learning method (and playing style) is very different from *Deep Blue*'s, and is a harbinger of the future of "narrow AI." *AlphaGo* "learns" via a two-pronged deep-learning approach that uses "value networks" to evaluate board positions and "policy networks" to select moves.[217] The DLNNs are trained partly by supervised learning using a dataset of human expert games (with about 30 million total moves), and partly from unsupervised reinforcement learning from games that it played with itself (using Monte Carlo tree search programs to simulate many thousands of random games). Notably, *AlphaGo* does not use look-ahead search; instead, its moves are a consequence of a gestalt-like holistic assessment of a single "Go position" (as one "pattern"). Where *Deep Blue* is essentially an expert system built using handcrafted rules, *AlphaGo* uses general machine-learning techniques to effectively

[215] M. Campbell, "Knowledge discovery in Deep Blue," *Comm. of the ACM* 42, Nov. 1999. Essentially the same "brute force" learning/playing method was used about 10 years later by a widely available commercial chess program, *Deep Fritz*, to defeat the then-reigning chess champion Vladimir Kramnik, but in which the AI system was run on a personal computer, evaluated only 8 million positions per second, and searched to an average depth of 17 to 18 moves (Ref: http://news.bbc.co.uk/2/hi/europe/6212076.stm). Compare this to the hardware required by *Deep Blue*, a 30 node, massively parallel system enhanced with 480 special purpose chess chips. When *Deep Blue* defeated Kasparov, its underlying hardware was the 259th most powerful supercomputer in the world (M. Newborn, *Deep Blue*, Springer-Verlag, 2002).

[216] The chess-playing program *Pocket Fritz 4* reached grandmaster level in 2009 while running on a 528 MHz HTC Touch HD mobile phone: https://en.chessbase.com/post/pocket-fritz-4-full-ahead-.

[217] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature* 529, 28 Jan 2016.

teach itself a complex "board position → move" mapping function, f_{Go}: *position* → *move*. The caveat of *AlphaGo*'s human+ level playing ability—and an example of the "harbinger of future AI" aspect of its landmark achievement—is that short of watching what move *AlphaGo* selects for a given board position, the details of the function f_{Go} are effectively invisible. While this "unknowability" caveat is nothing new—it is a well-known characteristic of all neural-net based learning[218]—its appearance in future military AI-based weapon systems is all-but-guaranteed if similar learning methods are used, and raises the more ominous spectre of military autonomous systems sometimes behaving unpredictably (a subject we will revisit in a later section).

In the case of *Deep Blue* and *AlphaGo*, both of their human victims were surprised by their AI opponent at some point during their respective matches: Kasparov, by a human-move-like sacrifice of a pawn in the first time (though the surprising move was later revealed to be a result of a programming error[219]); and Lee SeDol, by a move that was so surprising—"not a human move"—that the human player had to leave the room for 15 minutes to recover his composure.[220]

There is one other recent landmark achievement—*Turing Learning*[221]—that, though it does not yet surpass human performance (indeed, it is the first of its kind), nonetheless highlights the rapidly growing power of machine learning techniques.

Turing Learning is the first "self-learning" AI-system (based on a "simple" single-hidden-layer neural network architecture) that is able to effectively *infer the rules that govern the behavior of individual robots within a robotic swarm by watching the swarm*. Moreover: (1) collective behaviors can be directly inferred from motion trajectories of a single agent in the swarm, and (2) the basic technique can be applied to any observed system, human or machine. It works by simultaneously optimizing two populations of computer programs: one population represents models of the

[218] M Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, 2003.

[219] During an interview after the match, Kasparov was so "stunned" by *Deep Blue*'s seeming sacrifice of a pawn in the first game—"a wonderful and extremely human move"—that it altered the way he played in subsequent games, and, arguably, contributed to his eventual defeat. Ref: G. Kasparov, "The day that I sensed a new kind of intelligence," *Time*, 25 March, 1996). However, 15 years later, one of *Deep Blue*'s designers suggested that the move was due to a programming bug: K. Finley, "Did a computer bug help Deep Blue beat Kasparov?", *Wired*, 28 Sep 2012.

[220] C. Metz, "The sadness and beauty of watching Google's AI play Go," *Wired*, 11 March, 2016.

[221] G. Templeton, "Turing Learning breakthrough: Computers can now learn from pure observation ," *ExtremeTech*, 30 Aug 2016: https://www.extremetech.com/extreme/234669-turing-learning-breakthrough-computers-can-now-learn-from-pure-observation.

behavior of the system under investigation, and the other represents the classifiers. Two robot swarms are used: "A" (the "true" swarm) and "B" (the "learning" swarm). The movements of both "A" and "B" are tracked by the two learning systems: the classifier is rewarded for its ability to discriminate between "A" and "B"; and the model is rewarded for its ability to *fool* the classifier. Unlike other system identification methods, Turing Learning does not require any predefined metrics to quantify the difference between observed behaviors of a system and its models; indeed, Turing Learning outperforms metric-based system identification methods, in terms of model accuracy.[222] Although it is impossible to predict how this prototype technology might evolve in the coming years, it is safe to say that its continued development may impact the operational deployment options of future autonomous weapon systems (e.g., robotic swarms may be endowed with the ability to adapt their behaviors in real-time by observing enemy swarms).

Where state-of-the-art AI still falls short

Despite the string of recent successes in "narrow AI" (in which an AI system is taught, or learns, how to perform on a specific class of problems), there are many problems for which "narrow AI" techniques still fall far short of "solving." And even among those problems for which "narrow AI" both is well suited and demonstrably outperforms humans (such as it does for any of the problems listed in table 2), there are situations when an AI's "solution" is *surprising* and/or blatantly *wrong*.

While we may certainly expect to be "surprised" by an AI system's "solution" to a hard problem (such as Lee SeDol's surprise at one of *AlphaGo*'s moves during the second game of the landmark match he lost in 2016),[223] a limitation that applies to *all* extant machine learning methods as they apply to "narrow AI" problems is that they are effectively "black boxes" that do not easily reveal the "logic" behind the "reasoning."[224] This may be innocuous when playing an AI system in chess, but it assumes an entirely new (and serious) dimension if the "narrow AI" in question is embedded within a military autonomous system. For example, how does one ensure

[222] W. Li, M. Gauci, and R. Gross, "Turing learning: a metric-free approach to inferring behavior and its application to swarms," *Swarm Intelligence*, Vol. 10, No. 3, September 2016: http://link.springer.com/content/pdf/10.1007%2Fs11721-016-0126-1.pdf.

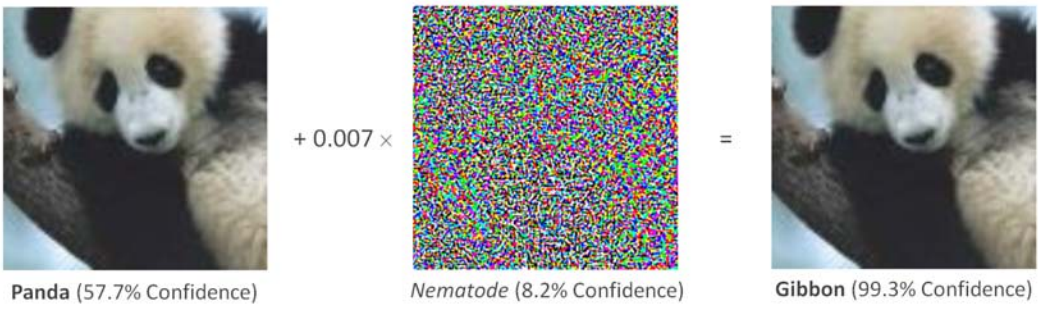
[223] C. Metz, "The Sadness and Beauty of Watching Google's AI Play Go," *Wired*, 3 March 2016: http://www.wired.com/2016/03/sadness-beauty-watching-googles-ai-play-go/.

[224] In the context of AI-based text-processing systems, MIT has recently introduced a method to train neural networks so that they provide rationales for their otherwise (and traditionally) opaque classifications. Ref: T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing Neural Predictions," Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, 2016. https://people.csail.mit.edu/taolei/papers/emnlp16_rationale.pdf.

(during, say, the testing and evaluation phase of DoD's acquisition process) that the autonomous system being developed will not perform "surprising" (i.e., unanticipated) actions during a mission? (A discussion of the issues involved in answering this question appears later in the narrative.)

The second issue—at least as egregious as displaying impenetrably surprising behaviors—is that otherwise well-performing "narrow AI" systems can also sometimes (and unpredictably) provide *bad* solutions to problems, with counter-intuitive properties. For example, two recent studies of state-of-the-art visual classifiers show that: (1) changing an image that has already been correctly classified in a way that is *imperceptible to humans* can cause a deep-learning neural network (DLNNs) to classify the image as something entirely different,[225] and, conversely, (2) it is easy to produce images that are *completely unrecognizable to humans* but that are "classified" by state-of-the-art DLNNs with 99.99% confidence (e.g., labeling with certainty that white noise static is a lion).[226] Figure 13 shows an example of the first case, in which a correctly classified image of a panda (with 57.7% confidence) is combined with a random image—which the classifier algorithm designates with low confidence as an image of a nematode—to produce an image that is now incorrectly, and with high confidence, classified as a gibbon, even though the first and third images are indistinguishable to a human.

Figure 13. Example of a DLNN's "blind spot" in recognizing images (see text)



Ref: After figure 1 in I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations* (ICLR) 2015: https://arxiv.org/pdf/1412.6572v3.pdf.

[225] C. Szegedy et al., "Intriguing properties of neural networks," presented at the International Conference on Learning Representations, 2014: https://arxiv.org/abs/1312.6199.

[226] A. Nguyen, J. Yosinski, and J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *Computer Vision and Pattern Recognition* (CVPR), IEEE, 2015: http://www.evolvingai.org/files/DNNsEasilyFooled_cvpr15.pdf. The authors of this study believe that *all* AI techniques that derive from creating decision boundaries between classes (not just deep neural networks) are subject to this "self fooling" phenomenon.

To emphasize: All deep learning neural networks effectively have "blind spots" in the sense that their input space inevitably contains elements that are arbitrarily close to correctly classified examples but that are misclassified.[227] Moreover, these "blind spots" display a kind of universality since the same misclassifications typically appear both in different DLNN architectures trained on the same dataset and by the networks trained on different data (i.e., misclassifications are not *just* a consequence of overfitting to a particular model). In a military context, such "blind spots" represent both a *vulnerability*—to a novel form of cyber intrusion by an adversary, whereby just the right array of pixels is injected into, say, the DLNN-trained image sensors of an autonomous system to render its environment temporarily unrecognizable)—and a *weapon*, whereby friendly forces do the same to an adversary's autonomous systems.

Finally, and before moving on to the "general AI" problem, note that all of the "narrow AI" successes thus far cited share two fundamental characteristics (apart from the obvious fact that they are all designed to "solve" specific problems—e.g., play a good game of checkers, or chess, or Go):

1. *They map fairly simple inputs to outputs.* For example, an image (as input) is classified as, say, a "dog" (as output) by an image recognition program; the sentence "this phrase is in English" (as input) is translated to, say, its Russian equivalent (as output) by an AI-translator algorithm; or, as an example of an ostensibly "more complex" variant, the moving 3D video (as input from, say, a self-driving car's cameras and other sensors) is "transformed" (via the "narrow AI" system) into a new position/movement-vector for the car.

2. *The times scales for human performance (on the same set of specific problems) are fairly short.* Whether a typical human is "good" at solving a specific problem or merely adequate, if the problem is such that the human processing time is on the order of *seconds*, today's state-of-the-art AI can probably automate (if not exceed a typical human's ability to perform) the specific task. (This is not to say that games such as chess or Go can be "solved" in a few seconds—only that each essentially requires but a single "glance" at the board position to provide the information necessary for making a move.)

One other common element is that all of these systems require huge datasets for training.[228] For example, the image classifiers that competed in the 2014 Large Scale

[227] M. James, "The Flaw Lurking In Every Deep Neural Net," *I Programmer*, 27 May 2014.

[228] In contrast, most humans are capable of "one shot" learning, whereby general categories can be self-learned "on the fly" using only a few examples. This is an (as yet, unavailable) ability that is desirable to have embedded in the software driving autonomous systems and that

Visual Recognition Challenge (LSVRC)[229] all trained on a set of images distributed among 1,000 categories and 1.2 million images;[230] and training required significant human effort to provide a large enough sample space of "correct" labels.[231] Two fundamental problems with requiring large datasets for training are: (1) the spectre of hidden or otherwise latent patterns that may bias the data (and thereby inadvertently skew what AI systems "learn"), and (2) the difficulty and cost of simply acquiring the data (an issue that immediately rises to the forefront in any discussion of DNN-driven DoD procurement of AI systems). In the first case, it is not always clear whether what AI systems "learn" from a given dataset are bona fide patterns in the data or merely "half-truths" distorted by hidden or latent biases in the training data. For example, Microsoft recently unveiled an AI-driven chat bot called *Tay*[232] and almost immediately shut it down[233] because of concerns over its inability to recognize when it was making offensive or racist statements. While the bot was not "taught" to make racist comments, the developers failed to take into account the statistical likelihood that the bot would "learn" *all ambient social norms* prevalent in the data-rich tweet/chat space it is programmed to interact with. In the second case, since DoD has only recently embarked on a path toward developing and acquiring

comes with its own associated risks (and which therefore represents a bona-fide analysis "gap"). Ref: Google has recently announced a possible breakthrough in "one view" learning, exploiting the same memory-enhanced architecture used by *DeepMind* to learn to traverse the London Underground (see discussion in main text). When used to learn to recognize images, the system still needs to effectively bootstrap itself to learn several hundred categories, but after an initial "pre learning" phase, it is able to recognize new objects from a single image. Ref: O. Vinyals et al., "Matching Networks for One Shot Learning," *arXiv*, 1606.04080v1, 13 June 2016.

[229] LSVRC homepage: http://www.image-net.org/challenges/LSVRC/2014/index#introduction.

[230] http://image-net.org/challenges/LSVRC/2014/browse-synsets. These training data are a subset of the *Imagenet* database, which is a vast archive of more than 14 million images that have been identified by humans: http://image-net.org/.

[231] Google has developed a method that uses unsupervised learning to teach an AI system to distinguish among objects observed in YouTube videos. Using 1,000 computers (with 16K cores) to sort through 10 million 200×200 pixel images, the system learned to recognize 22K object categories with a 15.8% accuracy (with represents a 70% relative improvementover the previous state-of-the-art). Ref: Q. Le et al., "Building high-level features using large scale unsupervised learning," IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6639343.

[232] S. Perez, "Microsoft's new AI-powered bot Tay answers your tweets and chats on GroupMe and Kik," *TechCrunch*, 23 March 2016: https://techcrunch.com/2016/03/23/microsofts-new-ai-powered-bot-tay-answers-your-tweets-and-chats-on-groupme-and-kik/.

[233] Peter Lee, Corporate Vice President Microsoft Research, "Learning from Tay's introduction," Official Microsoft Blog, 25 March 2016: ttps://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.000805batb8df0w106j2gxbiibbsg.

autonomous systems—systems that will be tasked with performing in vastly different environments from those for which the commercial and academic research communities have built their own AI systems—it will be an enormous undertaking for DoD to build the requisite types of training datasets.

General AI and the ability to reason

One major area in which state-of-the-art AI currently falls short is in *general* AI, the basic challenge of which is to develop a system that both performs well across a wide spectrum of cognitive skills and responds to many different forms of input data. General AI directly impacts each of the core AI-related components of autonomous systems: *perception*, *navigation*, *planning*, *behavior*, *targeting*, and the *human-system interface*. For example, the scene-recognition part of "perception" may be adequately "solved" via "narrow AI" training methods for sufficiently well characterized environments; however, with an increase in the difference between the real dynamic operational environment and the ones depicted in the dataset used for training, the autonomous system must be able to move gracefully away from simple "recognition" to *reasoning*. And even rudimentary reasoning (constrained to "simple," well-defined unchanging environments—something that is unlikely to occur in real-world scenarios) requires an ability to *infer* properties in the world from sensor-derived data and past experience. Except for the simplest of environments, an autonomous system's *behavior* and *targeting* algorithms must be able to respond quickly to all elements in a constantly changing environment. While some of these elements may be foreseen (during training and programming) and others (for which the training may nonetheless be adequate), may not be foreseen, many will be genuinely unanticipated (in training) and therefore require "on the fly" reasoning. However, state-of-the-art AI-based inference generally lags far behind the abilities of the best "narrow AI" systems.

To be more precise, while computers surpass humans in their ability to reason *deductively*, they are currently far behind in their ability to perform either *inductive* or *abductive* reasoning. *Deductive* reasoning consists of applying standard rules of logic to known facts and forming true propositions that are *entailed* by those facts. For example, from the facts "all oranges are fruits" and "all fruits grow on trees," one can deduce that "all oranges grow on trees." In deductive reasoning, if something is true about a general class of things, it is true for all members of that class. As discussed earlier, expert systems are textbook examples of applying this approach. *Inductive* reasoning is the opposite of deductive reasoning, in that it consists of generalizing from specific observations. One makes many observations, discerns a pattern, generalizes to a description of what makes up the pattern, and infers an explanation or rule for the conditions under which given patterns will occur. All neural-net-based deep-learning systems are effectively inductive inference machines. And, as table 2 showed, the current generation of AI systems can already outperform

humans on many complex "limited domain" problems, although they still fall short on more open-ended problem domains.

Abductive reasoning,[234] on the other hand, is an inference of some fact X, for which there is no direct evidence (of the form, say, "if A then X," and "A is known to be true"), but there are indirect pieces of evidence such that, when they are combined, collectively "strongly suggest" that X is true. Although this reasoning process may appear "simple" to humans, it is far from trivial to implement in an AI system. Indeed, it has been argued that since local syntactical systems alone cannot be used to model abductive inference, no computational processes can produce true intelligence.[235] However, since it is not necessarily "true intelligence" that DoD needs to develop for its autonomous systems, but only those aspects of it necessary for the system to successfully execute its missions, it is instructive to investigate any recent AI systems that have been taught to use abductive reasoning. Edbia[236] provides a cogent critique of AI from the perspective of cognitive science.

Watson

A well-known example of an AI system that relies heavily on abductive reasoning, and that has received much recent publicity, is IBM's *Watson*;[237] (or, more precisely, the DeepQA[238] architecture that underlies *Watson*). *Watson* is part of a more general ongoing research effort that started in 2007. The goal of its development was to develop an AI system that performs sufficiently well on open-domain (free-form based) question-and-answering to compete with human champions at the game of

[234] "Abduction," *Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/abduction/.

[235] Fodor, J., *The Mind Doesn't Work That Way*, MIT Press, Cambridge, MA, 2000. Two additional challenges to implementing a human-level intelligence are the "symbol grounding" problem (which refers to establishing a direct correspondence between an AI system's internal symbolic representation and external real-world entities, events, and relationships), and the "framing problem" (which refers to problem of predicting what changes in the world are incurred as a result of an AI system's action(s), and what stays the same. Searle (*The Rediscovery of the Mind*, MIT Press, Cambridge, MA, 1992) has argued that because local syntactical systems cannot perform symbol grounding, computational processes cannot be intelligent; while Pylyshyn (*The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Ablex, Norwood, NJ, 1987) argues that it is impossible to model the world entirely by logical propositions.

[236] H. Ekbia, *Artificial Dreams: The Quest for Non-Biological Intelligence*, Cambridge University Press, 2008

[237] http://www.ibm.com/watson/.

[238] https://www.research.ibm.com/deepqa/deepqa.shtml.

Jeopardy!.[239] We discuss some details of this project here to illustrate both the potential for developing AI systems that are able to reason abductively and the technical challenges that remain to be solved (and that therefore must also be part of DoD's calculus of what capabilities are likely—and not likely—to be available in the near future).

Watson's claim to fame came in 2011, when it defeated the two highest ranked *Jeopardy!* players of all time in a two game match played on 14 and 15 Feb 2011.[240] The game challenges three human contestants to answer natural language questions over a broad range of subject domains, with penalties for wrong answers. In preparation for *Watson*'s 2011 game with Ken Jennings (who won 74 straight games in 2004) and Brad Rutter (who defeated Jennings in a tournament in 2005), 100 test matches were first conducted against a variety of human players, with *Watson* winning 65% of them;[241] and a 15-question test round was played with Jennings and Rutter, which *Watson* won (during which none of the contestants answered any of the clues incorrectly). *Watson*'s final winning score was *triple* that of the second place winner (Jennings).[242]

Notably, it took roughly four years of research and development by a core team of 20 researchers[243] to build DeepQA (though the effort also made use of prior related work

[239] http://www.jeopardy.com/. Some sample *Jeopardy!* questions: (1) " On Sept. 1, 1715 Louis XIV died in this city, site of a fabulous palace he built" (ans: "What is Versailles?"); (2) "Pseudonym of labor activist & magazine namesake Mary Harris Jones" (ans: "What is Mother Jones?"); and (3) "Sakura cheese from Hokkaido is a soft cheese flavored with leaves from this fruit tree" (ans: "What is Cherry?") – Ref: "Sample 'Jeopardy!' questions," *Arizona State University*, 2010: https://asunews.asu.edu/20101103_jeopardyquestions.

[240] S. Baker, *Final Jeopardy: Man vs. Machine and the Quest to Know Everything*, Houghton Mifflin Harcourt, 2011.

[241] A. Sostek, "Human champs of 'Jeopardy!' vs. Watson the IBM computer: a close match," *Pittsburgh Post Gazette,* 13 Feb 2011: http://www.post-gazette.com/tv-radio/2011/02/13/Human-champs-of-Jeopardy-vs-Watson-the-IBM-computer-a-close-match/stories/201102130241.

[242] Famously, IBM|*Watson* gave a single (albeit inexplicably) erroneous response to the clue (in the category, "U.S. Cities"): "Its largest airport was named for a World War II hero; its second largest, for a World War II battle," with the answer, "What is Toronto?" The correct response was "What is Chicago?" It was given by both Jennings and Rutter, and, as later revealed,was also IBM|*Watson*'s second choice. IBM|*Watson*'s lead developer (David Ferrucci) has suggested that the error occurred because "U.S. city" does not appear in the actual clue, and the underlying logic assumed that categories do not strongly correlate with the type of response.

[243] http://researcher.watson.ibm.com/researcher/view_group_pubs.php?grp=2099

conducted at IBM and elsewhere) and to bring the system's performance capability to "human expert" level.[244]

Although IBM's *Watson* research team is no longer focused on *Jeopardy!*, development work continues on applying DeepQA to other problems. Spurring this effort, is the fact that DeepQA's core algorithms for information retrieval, data mining free-form text documents, and associating "answers" to specific queries using statistical classifiers are not anchored to any particular subject-matter domain or even type of problem.[245]

As a concrete example of the steps required to adapt *Watson* to new domains, Ferrucci et al.[246] describe post-*Jeopardy!* work on applying DeepQA to health care—specifically, on performing *differential medical diagnoses*. The general problem is to develop a diagnostic support tool that uses the context of an individual medical case (as extracted from datasets that describe a patient's medical condition—e.g., a patient's electronic medical record) to generate a ranked list of candidate diagnoses with associated confidence levels. Various dimensions of evidence are weighed just as in IBM|*Watson*, but individual bits of data are now domain-specific elements such as symptoms, patient and family history, and current medications. When the confidence in a given hypothesis (i.e., a medical condition) exceeds a certain threshold, then, analogously to IBM|*Watson*'s "buzzing" the answer, "DeepQA|*Diagnostician*" returns the medical diagnosis.

Just as *Watson* uses abductive reasoning to *Jeopardy!*, DeepQA|*Diagnostician* uses abduction to discover a piece of structured or unstructured knowledge that patients with disease D show symptom S. If it is then uncovered that the patient in question has symptom S, the system will generate the hypothesis that the patient has disease D, and then systematically search for evidence to either support or refute this hypothesis. More complex cases of abduction are also possible. While the use of abduction in medical diagnoses has been proposed before, DeepQA is unique in its ability to generate alternative hypotheses, along with associated confidence levels and supporting evidence.

[244] "Human expert" level is achieved when the contestant can answer roughly 70% of the questions being asked with at least 85% precision in 3 seconds or less: http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=2160.

[245] J. Murdock and G. Tesauro, "Statistical Approaches to Question Answering in Watson," Mathematics Awareness Month theme essay, *Joint Policy Board for Mathematics*, 2012.

[246] D. Ferrucci, et.al., "Watson: Beyond Jeopardy!," *Artificial Intelligence* 199/200, 2013.

Ferrucci, et.al.[247] cite two main challenges in adapting DeepQA to act as a clinical decision support tool: (1) the challenge of embedding the tool into an existing clinical decision system (not unlike the challenge that a similar research effort would face in embedding a DeepQA-like technology into the existing document declassification process), and (2) the challenge of adapting the internal components of DeepQA itself to the medical domain.

The first challenge is the ability to tailor DeepQA's embedded NLP algorithms to extract clinically relevant information (from, say, EMR systems). The problem is nontrivial because different problem areas require different—domain-specific—information, and the nature of this information must be characterized well enough to tune the NLP algorithms to find and extract it. In the context of declassification, and apart from performing simple-minded "key word" searches, the first order of business is to understand *why* (or *why not*) a particular piece of information is "significant" in weighing evidence pro/con as to, say, a given document's classification level. Another component of this first kind of challenge is to tailor the output so that each hypothesis is associated with its supporting evidence in a manner consistent with existing practices of (in this example) medical practitioners. The authors acknowledge that "significant areas of research remain within the natural language processing capability of DeepQA itself."[248]

The second challenge involves adapting DeepQA to a specific domain. The authors identify three kinds of adaptations that must be made: (1) *content adaptation*, which refers to organizing the domain content for hypothesis and evidence generation; (2) *training adaptation*, which refers to the development of domain-specific sets of training question-answer pairs in order for the system to learn how to weigh individual bits of evidence; and (3) *functional adaptation*, which consists of adding new domain-specific question analysis, candidate generation, and hypothesis scoring algorithms. While none of these challenges preclude DeepQA-based applications to be developed for specific domains, they are important reminders of the conceptual and practical impediments to development time that any research effort must face when adapting an existing general-purpose DeepQA-like inference engine to a specific domain (such as a military autonomous system).

Examples of AI / "system" *failures*

Along with gaps in ability (as outlined in the "narrow-AI vs. general-AI" discussion above), AI can also *fail*, sometimes badly. While the details of specific failures depend

[247] Ibid.

[248] Page 100: D. Ferrucci, et.al., "Watson: Beyond Jeopardy!," *Artif. Intel.*, Vol. 199/200, 2013.

on the AI system (and the context in which a failure occurs), essentially all AI systems above a threshold level of "complexity" (read: those underlying the development of autonomous weapon systems) are prone to "surprising" behaviors—behaviors that are, at best, "unexpected" but inconsequential, and, at worst, both unexpected and catastrophic. We have already mentioned a few *academic* "failures," such as the "blind spot" that all deep-learning NN-based image classification systems suffer from. However, the impact of these kinds of failures, were they to occur in a military operational context, is—as yet—impossible to predict. But there are also historical examples of failures that have occurred in otherwise fully operational systems (albeit only loosely AI-like, and still not in military contexts) that collectively provide both a cautionary flag and "lessons learned" for developers of autonomous weapons systems:

- 1982 — *Software designed to make discoveries, discovers how to cheat instead.*[249] EURISKO (developed by Lenat, who later developed CyC (see first entry in table 2) was an early self-improving AI system that had built-in heuristics for suggesting new heuristics. During one of EURISKO's runs, it was found that the rank of a particular newly "discovered" heuristic kept increasing, though the heuristic appeared to serve no useful purpose. Upon closer inspection, it was discovered that EURISCO had inserted its own "name" (as "creator") beside an early-generation heuristic, and thereafter kept adding to its innately meaningless rank. In short, the program had found a way to cheat.

- 1983 — *Nuclear attack early warning system falsely claims that an attack is taking place.*[250] In a well-known incident that occurred on 26 September 1983, and during which a nuclear war was averted by human intervention, the Soviet Union's nuclear early warning system twice reported the launch of Minuteman intercontinental ballistic missiles from bases in the United States. Duty officer Lieutenant Colonel Stanislav Petrov correctly identified the warning as a false alarm.

- 1988 — *U.S.S. Vincennes shoots down Iranian passenger jet.*[251] In another well-known incident from the 1980s, the guided missile cruiser

[249] G. Johnson, "Eurisko, The Computer With A Mind Of Its Own," *The APF Reporter*, 1984: http://www.aliciapatterson.org/APF0704/Johnson/Johnson.html.

[250] A. Libak, "Nuclear War: Minuteman," *Weekendavisen*, 2 April 2004: http://www.brightstarsound.com/world_hero/weekendavisen.html.

[251] Lt. Col. (U.S. Marine Corps, Retired) D. Evans, "Vincennes: A Case Study," *Proceedings Magazine*, U.S. Naval Institute, Aug 1993: http://www.usni.org/magazines/proceedings/1993-08/vincennes-case-study.

USS *Vincennes* shot down an Iranian passenger jet in the Persian Gulf after the ship's Aegis targeting system erroneously identified it as a military fighter. While (in hindsight) the crew of the *Vincennes* had the necessary data to determine that the radar contact was a civilian aircraft, Aegis had, unfortunately, been designed to detect large Soviet bombers, not passenger jets. One scholar explains, "Even though the hard data was telling the crew that the plane wasn't a fighter jet, they trusted what the computer was telling them more. Aegis was on semiautomatic mode, but not one of the eighteen sailors and officers on the command crew was willing to challenge the computer's wisdom. They authorized it to fire."[252]

- 2010 — *Stock trading software causes a trillion dollar flash crash.*[253] The historic stock market crash that occurred on May 6, 2010, was caused by a large investor using automated trading software to sell futures contracts (used by traders to bet on the future performance of stocks in the S&P 500 index). The software sold a very large number of contracts very quickly, and inadvertently initiated a feedback loop whereby it reacted to the rise in trading volume it created by increasing the number of sell orders. In turn, a lack of buyers, coupled with the rapid selling of futures contracts, affected the underlying stocks and the broader stock indices, resulting in a runaway cascade failure of the market..

- 2015 — *A robot grabs and kills a man.*[254] An industrial robot at a Volkswagen production plant in Germany, designed to grab and configure auto parts, instead grabbed a factory worker and pushed him against a metal plate, crushing him.

- 2015 — *Image tagging software classifies black people as gorillas.*[255] Until "corrected," Google's *Photos* application [256] contained a fault whereby black people were classified as "gorillas." A similar "fault" was detected earlier in

[252] P. W. Singer, *Wired for War*, Penguin Books, 2009

[253] B. Rooney, "Trading program sparked May 'flash crash'," *CNN Money*, 1 Oct 2010: http://money.cnn.com/2010/10/01/markets/SEC_CFTC_flash_crash/.

[254] E. Dockterman, "Robot Kills Man at Volkswagen Plant," *Time*, 1 July 2015.

[255] T. Finley, "Google Apologizes For Tagging Photos Of Black People As 'Gorillas'," *The Huffington Post*, 2 July 2015.

[256] https://photos.google.com/.

the same year, when Google *Maps* was discovered to link to the address of the White House when queried with a racial slur.[257]

- 2016 — *Car autopilot navigation system kills driver.*[258] A *Tesla* Model S (a full-size all-electric automobile produced by Tesla Motors, and equipped with *Autopilot*, which allows limited hands-free driving)[259] was involved in a fatal crash on 7 May 2016. The driver was killed by an 18-wheel tractor-trailer, as it drove across the highway perpendicular to the car. Neither the driver nor the car detected the tractor-trailer "against a brightly lit sky" and brakes were not applied. This is the first known fatality in a *Tesla* where *Autopilot* was active.

[257] J. Crave, "If You Type 'N—— House' Into Google Maps, It Will Take You To The White House," *The Huffington Post*, 20 May 2015.

[258] J. Golson, "Tesla driver killed in crash with Autopilot active, NHTSA investigating," *The Verge,* 30 June 2016.

[259] https://en.wikipedia.org/wiki/Tesla_Model_S.

Complex Adaptive Systems

> The musical notes are only five in number but their melodies are so numerous that one cannot hear them all. The primary colors are only five in number but their combinations are so infinite that one cannot visualize them all. The flavors are only five in number but their blends are so various that one cannot taste them all. In battle there are only the normal and extraordinary forces, but their combinations are limitless; none can comprehend them all. For these two forces are mutually reproductive; their mutual interaction as endless as that of interlocked rings. Who can determine where one ends and the other begins?
>
> — Sun Tzu, *The Art of War*

Complex adaptive systems (CASs)[260] consist of (typically many) interconnected, nonlinearly interacting parts. Moreover, their aggregate behavior is emergent. That is to say, the properties of the whole are not possessed by, and are not directly derivable from, any of the parts; a water molecule is not a vortex, and a neuron is not conscious. A complex system must therefore be understood not just by listing the set of components out of which it is constructed, but by knowing the topology of interconnections, by knowing the interactions among those components, and—most importantly—by observing how it evolves over time and under different conditions. Figure 14 shows some examples of CASs.

Understanding the basic properties of CAS—and the methods and tools used to study them—can help shed light on what otherwise would be difficult-to-understand interrelationships among the operational benefits, challenges, costs, risks, and capabilities of unmanned autonomous systems. CASs and the tools used to study them pervade many key aspects of AI, autonomy, robotics, swarms, and human-machine cooperative systems (albeit not always in the most obvious fashion). For example, the power of deep-learning neural-networks, which lie at the heart of many

[260] S. Kauffman, *Investigations*, Oxford University Press, 2000; S. Wolfram, *A New Kind of Science*, Wolfram Media, 2002; J. Miller and S. Page, *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*, Princeton University Press, 2007; C. Mainzer, *Thinking in Complexity: The Computational Dynamics of Matter, Mind, and Mankind*, Springer-Verlag, 2007; M. Mitchell, *Complexity: A Guided Tour*, Oxford University Press, 2009; C. Gross, *Complex and Adaptive Dynamical Systems: A Primer*, Springer-Verlag, 2009.

of today's most sophisticated "narrow AI" systems, derives, in part, from the complex network of connections among the neurons distributed across various layers. This behavior falls under the broader rubric of dynamic network theory, which is itself a subfield of complex systems.[261]

Figure 14. Examples of complex adaptive systems

Ref: G. Mobus and M. Kalton, *Principles of Systems Science*, Springer-Verlag, 2015

Most modern algorithms used to define and control robots derive from complex-systems-theoretic behavior-based architectures.[262] And swarms are, of course, prototypical complex systems, and the only proper way to design and study their behavior is by evolutionary programming techniques[263] and multi-agent-based modeling.[264] (It is also worth mentioning that the general study of "autonomous

[261] A.-L. Barabasi, *Network Science*, Cambridge University Press, 2016.

[262] R. Brooks, *Cambrian Intelligence: The Early History of the New AI*, MIT Press, 1999.

[263] R. Eberhart, Y. Shi, and J. Kennedy, *Swarm Intelligence*, Morgan Kaufmann, 2001.

[264] H. Iba, *Agent-Based Modeling and Simulation with Swarm*, CRC Press, 2013.

systems" dates back to the rise of cybernetics and early systems theory in the 1960s, a precursor of modern-day complex systems theory.)[265]

We will exploit this close connection between complexity theory and the main subject of this report (i.e., AI, robotics, and swarms) in three ways: (1) by deepening an understanding of the basic properties of *individual* autonomous weapon systems (since their AI behavior derives from complex systems theoretic architectures), (2) by deepening an understanding of *robotic swarms* (since collective behavior is an emergent self-organized complex process), and (3) by ameliorating some of the Testing and Evaluation (T&E) and Verification, Validation, & Accreditation (VV&A) challenges that autonomous weapon systems will inevitably face within DoD's existing acquisition process. Many of the same technical challenges now facing the design and development of unmanned autonomous systems (UASs) have been dealt with, with varying degrees of success over the last several decades, within the complex systems research and modeling communities. While a few of these "lessons learned" can, in principle, be applied almost immediately to the design of (and deeper understanding of the general behavior of) autonomous systems, others are best viewed as cautionary notes, and describe the inherent limitations of using various forms of analysis to probe into a CASs' behavior.

Complexity theory also provides valuable insights into how risk and vulnerability may be assessed for autonomous systems, including human-machine hybrids. The first step is to understand how complex systems can *fail*.

Basic properties

Gases, fluids, crystals, and lasers are all familiar kinds of complex systems in the realm of physics. Chemical reactions, in which a large number of molecules conspire to produce new molecules, are also good examples. In biology, there are DNA molecules built up from amino acids, cells built from molecules, and organisms built from cells. On a larger scale, the national and global economies and human culture as a whole are also complex systems, exhibiting their own brand of global cooperative behavior. One of the most far-reaching ideas of this sort is James Lovelock's controversial *Gaia hypothesis*,[266] which asserts that the entire earth—its oceans, rocks, atmosphere, and entire biosphere—is essentially one huge, complex organism, delicately balanced on the edge-of-chaos. Perhaps the quintessential example of a complex system is the human brain, which, consisting of something on

[265] T. Rid, *Rise of the Machines: A Cybernetic History*, W. W. Norton, 2016.

[266] J. E. Lovelock, *Gaia: A New Look at Life on Earth*, Oxford University Press, 2000.

the order of 10^{10} neurons with 10^3 to 10^4 connections per neuron, is arguably the most "complex" complex system on this planet. Somehow, the cooperative dynamics of this vast web of "interconnected and mutually interacting parts" manages to produce a coherent and complex enough structure for the brain to be able to investigate its own behavior.

Roughly speaking, the *science* of CAS emerged between the late 1980s and early 1990s,[267] and did so with a strongly interdisciplinary cast, involving the efforts of physicists, chemists, biologists, computer scientists, and social scientists. However, its conceptual origins date back to Alan Turing's observations of biological pattern formation in the early 1930s,[268] and the rise of systems theory and cybernetics in the 1950s.[269] Its association with social science is due largely to Axtell's and Epstein's *Sugarscape* model,[270] which used agents to model economic dynamics. Around the same time that *Sugarscape* was being developed (late 1990s), CNA introduced its pioneering CAS-based models of land combat (ISAAC and EINSTein).[271] Today, basic research on complex systems theory remains unabated and widespread. Two well known (but no means only) centers of study are the Santa Fe Institute (SFI)[272] and New England Complex Systems Institute (NECSI).[273]

Consider some examples of the dynamics of complex systems: predator-prey relationships in natural ecologies, dynamics of world financial markets, firing patterns of neurons in a human brain, information flow on the internet, antigen-antibody interactions in an immune system, pedestrian and vehicular traffic dynamics, the behavior of ant and termite colonies, the spread of infectious disease, the dynamics of combat, and (what the narrative of this report purports to show) unmanned autonomous systems.

---

[267] See, for example, the landmark series of workshop publications that collectively chronicle the rise of the Santa Fe Institute in Santa Fe, New Mexico: D. Stein, editor, *Lectures in the Sciences of Complexity*, Addison Wesley, 1989; E. Jen, editor, *1989 Lectures in Complex Systems*, Addison-Wesley, 1990; L. Nadel and D. Stein, *1990 Lectures in Complex Systems*, Addison-Wesley, 1991; G. Cowan, D. Pines, and D. Meltzer, *Complexity: Metaphors, Models, and Reality*, Addison-Wesley, 1994.

[268] A. M. Turing, "On computable numbers with an application to the Entscheidungsproblem," *Proceedings London Mathematical Society*, Vol. 42, 230-265, 1936.

[269] L. von Bertalanffy, *General System Theory*, Braziller, 1968; W. Ross Ashby, *An Introduction to Cybernetics*, Chapman & Hall, London, 1956.

[270] J. Epstein and R. Axtell, *Growing Artificial Societies: Social Science from the Bottom up*, MIT Press, 1996.

[271] A. Ilachinski, *Artificial War: Multiagent-Based Simulation of Combat*, World Scientific, 2004.

[272] https://www.santafe.edu/.

[273] http://necsi.edu/.

What all of these systems have in common is that they share (a significant number of) the following list of nine properties, described below (in no particular order):[274]

Many interconnected nonlinearly interacting heterogeneous parts

Complex systems owe their apparent complexity to the fact that they consist not just of isolated parts, but of deeply interconnected parts that continually respond to (and change as a function of) changes undergone by other parts to which they are connected. While the parts are usually related in some manner, and may all belong to the same general class of possible system constituents (for example, a predator-prey ecology may include multiple instances of the class "shark," and an urban traffic environment may include many different kinds of "automobile"), how one specific part interacts with another part may also, in general, be a function of the specific part and/or its previous history. How a hungry shark responds to prey in its immediate environment may be very different from how another, more satiated, shark responds.

The most interesting interactions are those that are nonlinear—i.e., those that entail disproportionate responses to (local) information. For example, the magnitude of a single neuron's electrical impulse does not steadily increase with increasing local chemical potential but is triggered, nonlinearly, in an all-or-none reaction once it senses a threshold local potential. Or, consider some site (the "part") on the world-wide web (the "system") that languishes for months, attracting few visitors, until, by chance, a word or phrase that appears on the site comes into vogue, is picked up by automated search-spiders, and is catalogued by web search engines. Suddenly, and in an unpredictably nonlinear fashion, the site is now besieged by visitors. In general, nonlinear interactions imply that small system perturbations may (though not necessarily must) cause a large effect.

Multiple simultaneous scales of resolution

Complex systems tend to be organized hierarchically, with complex behavior arising from the interaction among elements at different levels of the hierarchy. A biological organism is, simultaneously, the complex system comprising DNA, proteins, cells, tissues, and organs, as a whole. Similarly, weather consists of patterns on multiple scales, ranging from the individual molecules of the atmosphere, to small dust devils, to tornados, to full-blown hurricanes that span hundreds of miles.

[274] A. Ilachinski, *Cellular Automata*, World Scientific, 2001.

The individual parts of complex systems (which we will henceforth call, generically, agents) form groups that then effectively act as higher-level agents that, in turn, also cluster and interact with still other agents; these (still higher-level) groups, in turn, form super-groups than also act as agents, interacting with other agents (on different timescales); and so on. Koestler[275] observes that an agent on any given level of a complex system's hierarchy is driven by two opposite tendencies: (1) a integrative tendency, compelling it to function as a part of the larger whole (on higher levels of the hierarchy);, and (2) a self-assertive tendency, compelling it to preserve its individual autonomy.

Multiple metastable states

Complex systems generally harbor multiple metastable states. Multistable systems have multiple stable fixed points; which particular stable fixed point a system is attracted to depends on the initial configuration of the system. A metastable system is a system that is above its minimum-energy state, but requires an energy input if it is to reach a lower-energy state. Metastable states are thus states that are in a pseudo-equilibrium. Small perturbations to the system lead to recovery, but larger ones can also ignite large changes in the system. A metastable system can act as if it were stable, provided that all energy inputs remain below some threshold. Because one must keep track of multiple metastable states, the dynamics of such systems are often difficult to analyze mathematically, a task that is made harder still because it also usually involves dealing with local frustration (i.e., conflicting constraints that make it impossible to solve for globally minimal energy states).

The fact that complex systems almost certainly harbor multiple metastable states is important, on the conceptual level, because it compels the researcher to explore as large a volume of a system's (typically, very large) multidimensional state space as possible. A moment's thought will show that this deceptively simple assertion radically shifts the way in which complex systems are studied. It is commonly assumed that a system has but one, mathematically well-defined, equilibrium state. Once that state is "solved" for—either as a closed-form "solution" or by running a simulation—the system is said to be understood. However, if the system in question is a CAS, and harbors multiple metastable states—none of which are generally "solvable" in closed form, and the entire set of which depends on where a system "starts" its evolution—one can hope to characterize the "whole system" only by

[275] A. Koestler, *Janus: A Summing Up*, Random House, 1978. This book presages Kauffman's compelling edge of chaos hypothesis, articulated about a decade later: S. Kauffman and S. Johnson,"Coevolution to the edge of chaos," in C. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, editors, *Artificial Life II*, Addison-Wesley, 1992, pages 325-370.

exploring as many attainable states as possible. This therefore also implies that one must explore as many system evolutions as possible, which start from as many initial configurations as possible. A given state of a CAS can generally be understood only by providing a context for its being—i.e., by understanding all of the possible states of a system, and all of the possible ways these states can be attained.

Local information processing

The agents of a complex system typically "see" (and interact with) only a limited portion of the whole system, and act locally; i.e., interagent dynamics is usually strongly decentralized. There is no God-like "oracle" dictating what each and every agent ought to be doing; no master neuron tells each neuron of a brain when and how to "fire." Instead, the parts of the system act locally, using only local information. The order that emerges on a global scale does so naturally, and does not depend on either a central or an external control. Kauffman[276] observes that "contrary to our deepest intuitions, massively disordered systems can spontaneously 'crystallize' to a very high degree of order." Self-organization takes place as a system reacts and adapts to its external environment, with which it also usually has an open boundary (i.e., energy and/or other system resources are continually exchanged between the inside and outside of a system).

Self-organization

Self-organization is a fundamental characteristic of CAS. It refers to the emergence of macroscopic nonequilibrium organized structures, and is due to the collective interactions of the constituents of a complex system as they react and adapt to their environment. At first sight, self-organization appears to violate the second law of thermodynamics,[277] which asserts that the entropy S of an isolated system never decreases (or, more formally, $dS/dt \geq 0$). Since entropy is essentially a measure of the degree of disorder in a system, the second law is usually interpreted to mean that an isolated system will become increasingly more disordered with time. *How, then, can structure emerge after a system has had a chance to evolve?*

Upon closer examination, we see that self-organization in a complex system does not really violate the second law. The reason is that the second law requires a system to be *isolated*; that is, it must not exchange energy or matter with its environment. For nonisolated systems consisting of noninteracting or only weakly interacting particles,

[276] S. Kauffman, *Origins of Order*, Oxford University Press, 1993.

[277] F. Mandl, *Statistical Physics*, Wiley, 1988.

S consists of two components: (1) an internal component, S_i, due to the processes taking place within the system itself, and (2) an external component, S_e, due to the exchange of energy and matter between the system and the environment. The rate of change of S with time, dS/dt, now becomes $dS/dt = dS_i/dt + dS_e/dt$. As for an isolated system, $dS_i/dt \geq 0$. But there is no such constraint on dS_e/dt. If dS_e/dt is sufficiently less than zero, the overall entropy of the system can itself decrease. Thus, the entropy of a nonisolated system of noninteracting or only weakly interacting particles can decrease due to the exchange of energy and/or matter between the system and its environment.[278]

Emergent behavior

Emergence is one of the central ideas of complex systems theory. Emergence refers to properties of the whole system that are not possessed by, and are not directly derivable from, any of the system's parts. Or, more colloquially, emergence = *novelty*; i.e., complex systems will (almost always) *surprise us with their behavior*. A line of computer code cannot calculate a spreadsheet, an oxygen molecule is not a tornado and (unfortunately) no one can predict a significant gain (or catastrophic crash) of the stock market. Emergent behaviors, which appear on the macroscale, are typically novel and unanticipated, at least with regard to our ability to predict them from knowing a system's microscale parts and rules alone. Indeed, it is the microscale that induces the macroscale behavior. Some elements of emergent behaviors may be universal, in the sense that more than one set of local rules may induce more or less the same global behavior.

Examples of emergence include the characteristic spirals of the Belousov-Zhabotinski chemical reaction;[279] the Navier-Stokes-like macroscopic behavior of a lattice gas that

[278] The situation is more complicated for nonisolated systems consisting of strongly interacting particles and when the system is no longer in equilibrium with the environment. The second law effectively asserts only that a system tends to the maximum disorder possible, within the constraints due to the dynamics of the system (Y. Bar-Yam, *Dynamics Of Complex Systems*, Westview Press, 2003).

[279] The Belousov-Zhabotinski reaction is a chemical reaction consisting of simple organic molecules that is characterized by spectacular oscillating temporal and spatial patterns. One variant of the reaction involves the reaction of bromate ions with an organic substrate (typically malonic acid) in a sulfuric acid solution with cerium (or some other metal-ion catalyst). When this mixture is allowed to react exothermally at room temperature, interesting temporal and spatial oscillations (i.e., chemical waves) result. The system oscillates, changing from yellow to colorless and back to yellow about twice a minute, with the oscillations typically lasting for over an hour (until the organic substrate is exhausted): I. Motoike and A. Adamatzky, "Three valued logic gates in reaction-diffusion excitable media," *Chaos, Solitons & Fractals* 24, 2005, pp. 107-114.

consists, on the micro-scale, of simple unit-bit billiards moving back and forth between discrete nodes along discrete links;[280] and the seemingly purposeful task of forming clusters of randomly distributed objects—a behavior common in, say, ant colonies organizing the carcasses of their dead companions—that spontaneously and quite naturally emerges out a simple set of autonomous actions having nothing to do with clustering per se (as demonstrated by Beckers et al. in the context of exploring collective robotics[281]). The macroscopic behavior in each of these examples is unexpected, despite the fact that the details of the microscopic dynamics are well defined.

Nonequilibrium patterns and order

The long-term behavior of a complex system usually consists of a *nonequilibrium order*. This term refers to an organized state that remains stable for long periods of time despite matter and energy continually flowing in and out of the system. Nonequilibrium states are also sometimes called either *dissipative structures*[282] or *autopoietic systems*.[283] A vivid example of nonequilibrium order is the *Great Red Spot* on Jupiter (though any terrestrial hurricane will do just as well). This gigantic whirlpool of gases in Jupiter's upper atmosphere—which can fit three Earth-sized planets within its boundary—has persisted for a much longer time (at least 400 years) than the average amount of time any one particular gas molecule has spent within it. Despite the millions of individual molecules that have traveled in and out of the Great Red Spot—a substantial fraction of which have likely done so repeatedly, some perhaps also circumnavigating Jupiter's entire atmosphere—the Great Red Spot itself, as a high-level emergent entity, remains in a stable but nonequilibrium ordered state.

[280] J. Rivet and J. Boon, *Lattice Gas Hydrodynamics*, Cambridge University Press, 2005.

[281] R. Becker, E. Holland, and J. Deneubourg, "From local actions to global tasks: stigmergy and collective robotics," pages 181-189 in *Artificial Life IV*, edited by R. Brooks and P. Maes, MIT Press, 1994.

[282] I. Prigogine, *From Being to Becoming*, Freeman and Company, 1980.

[283] F. Varela, J. Humberto, R. Maturana, and R. Uribe, "Autopoiesis: the organization of living systems, its characterization and a model," *Biosystems* 5, 187-196, 1974.

Emphasis on process and adaptation rather than static structure

A CAS is almost never stagnant. It continually interacts with, and adapts to, changing conditions of its environment, and always evolves. It can neither be captured, conceptually or mathematically, nor be understood, by simply cataloging its parts and the rules according to which they interact. Such static "snapshots" never adequately capture the often latent and subtle patterns that such systems exhibit over long times. It is for this reason that computer simulations of complex systems are indispensable tools for studying them. Mathematical descriptions and/or equations of motion are often rendered tractable only if one makes a number of mean-field-like simplifications (such as assuming a strict homogeneity of parts and homogenous interactions); therefore, by themselves, they are rarely able to capture any emergent behaviors aside from the simplest. In fact, much of what falls under "complex systems research" consists not so much of "solving" equations, or of recording what state a given system is in at what time, as of patiently and systematically observing—and learning to recognize the properties of emergent patterns in—the behaviors that a system exhibits over the course of its (typically open-ended) evolution; and then repeating the process for many different starting conditions. Complex systems theory is, essentially, the art of finding the proper global context in which the local behavior can be understood.

The most interesting behavior is poised between chaos and order

Effective computation, such as that required by life processes and the maintenance of evolvability and adaptability in complex systems, requires both the storage and transmission of information. If correlations between separated sites (or agents) of a system are too small—as they are in the ordered regime shown in the schematic illustration on the previous page—the sites evolve essentially independently of one another and little or no transmission takes place. On the other hand, if the correlations are too strong—as they are in the chaotic regime—distant sites may cooperate so strongly so as to effectively mimic each other's behavior, or worse yet, whatever ordered behavior is present may be overwhelmed by random noise; this, too, is not conducive to effective computation. It is only within the phase transition region, in the complex regime poised at the edge-of-chaos, that information can propagate freely over long distances without appreciable decay. However loosely defined, the behavior of a system in this region is best described as complex—i.e., it neither locks into an ordered pattern nor dissolves into an apparent randomness. Systems poised at the edge-of-chaos are optimized, in some sense, to evolve, adapt, and process information about their environment: they are both stable enough to

store information, and dynamically amorphous enough to be able to successfully transmit it. [284]

Modeling lessons from studies of CAS

CASs are inherently difficult to study analytically—that is, by reductive methods that assume that the properties of a system may be deduced (and, implicitly, that the system itself may be understood) by decomposing it into progressively smaller and smaller parts. By analyzing a system in this way, there is a good chance that the most interesting properties of the system will become lost; the chance of this happening only increases as the complexity of a system's behavior increases. Understanding a complex system requires both analysis and synthesis. Think of the absurdity of seeking a clue to how consciousness arises by systematically stripping a brain down to a few neurons. In meticulously probing the parts, the analytical-reductionist method inevitably loses sight of the whole. The understanding of complex systems also requires a complementary holistic, or constructionist, approach, executed in parallel with reducing a system down to its essential parts, in which one explores how the system's parts cooperatively synthesize the whole. It is often the case that *qualitative* factors are the critical drivers (i.e., components of the system that involve human interaction or decision-making).

While no all-encompassing "formal theory" of complexity yet exists (in the sense that it does not yet constitute a rigorous set of mathematical "theorems" or universal "laws"), complexity is both mature enough as an multidisciplinary field of study and rich enough in its store of phenomenologically observed behaviors in a wide variety of physical systems, to offer deep insights into what modeling approaches are more (or less) suited for a given system, as well as general observations and lessons about the forms the most "useful" models should take.

The first, and key, property of many CASs is that they can display remarkably complex global, emergent behavior *despite* what may appear to be (sometimes, trivially) "simple" local rules.[285] To the extent that the basic principles of CASs also apply to the design (and behavior) of UASs, the main "takeaway" for system developers is that the behavior of UASs cannot, in general, be easily derived or predicted from the list of "local rules" alone; and that this general observation holds equally true for *individual* vehicles (which operate autonomously and not as part of a swarm, but whose own overall behavior emerges from lower-level rule-based

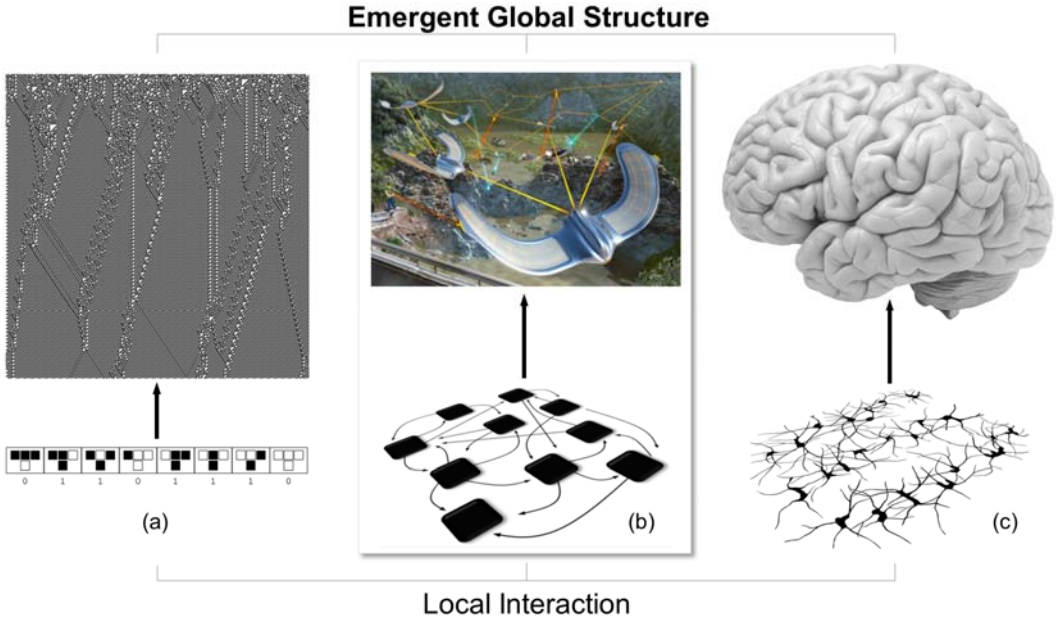
[284] S. Kauffman, *At Home in the Universe*, Oxford University Press, 1995.

[285] G. Mobus and M. Kalton, *Principles of Systems Science*, Springer-Verlag, 2015.

behaviors) and *swarms* (whose collective dynamics derives from rules of interaction among individual systems).

Figure 15 shows, schematically, three examples of how "simple" local interactions induce complex global structures: figure 15-a shows a pedagogical example of a one-dimensional cellular automaton (discussed below); figure 15-b shows an example of robotic swarms (discussed in a later section); and figure 15-c illustrates the emergence of a human brain—and, by implication, consciousness—from a substrate of individual neurons.[286]

Figure 15. A schematic illustration showing the ubiquitous emergence of complex global behavior from "simple" local interactions



Ref: U. Wilensky and W. Rand, *An Introduction to Agent-Based Modeling*, MIT Press, 2015.

Cellular automata

While there is obviously not enough space in this report to provide anything more than a cursory look at how difficult it is, in general, to predict global behaviors in complex systems by knowing only how their constituent parts interact, the salient points can be made by referring to figure 15-a. The bottom of figure 15-a (i.e., the

[286] C. Koch, *Consciousness*, MIT Press, 2012.

"local interaction") shows a particular rule for a one-dimensional two-state cellular automaton (CA); there are 256 possible rules in this "elementary rule space" (see below). CA were introduced by John von Neumann in the early 1950s as simple models of life in general, and biological self-reproduction in particular.[287] They have been continually studied over the years because they capture many essential features of complex self-organizing cooperative behavior observed in real systems.[288]

CA are easy to describe: imagine an infinite row of discrete sites (in practice, we are limited to some finite number: 600 sites are used to generate the top of figure 15-a), each of which harbors one of two values, 0 or 1 (which we can think of as being *off*, and denote as white, or *on*, and color black). A given sequence of site values, at time t, represents the global state of the CA at that time. The value of i at time t, $\sigma(i, t) \in \{0,1\}$, is a function of the values of site i and i's left and right neighbors at time $t-1$: $\sigma(i, t) = f[\sigma(i-1, t-1), \sigma(i, t-1), \sigma(i+1, t-1)]$. The function f may be defined as an explicit list of the values the center site assumes for all 8 possible 3-tuples that represent the values sites i, $i-1$, and $i+1$ have on the previous time step. The bottom of figure 15-a defines the rule-110:[289] (0,0,0) → 0, (0,0,1) → 1, (0,1,0) → 1, (0,1,1) → 1, (1,0,0) → 0, (1,0,1) → 1, (1,1,0) → 1, (1,1,1) → 0.

The top of figure 15-a shows a space-time plot (with time moving downwards) starting with a random initial string of 600 ones and zeros. In all, 600 time steps are depicted; each successive row represents the updated configuration of *on* and *off* states at the next time step. The obvious intricacy of the emergent global pattern, which may be viewed at two scales—on the individual site-level, by explicitly reading off the values of the individual cells, or on a higher level as propagating particle-like structures superimposed on a periodic background—belies an even deeper, unexpected, layer of complexity. Namely, this rule has been proven mathematically to be *universal*.[290] This means that with a proper selection of initial conditions (i.e., the initial distribution of *on* and *off* sites), rule-110 can be turned into a general purpose computer. The "deeper level" of complexity that this, in turn, implies is that the

[287] J. von Neumann, "The general and logical theory of automata," in *Cerebral Mechanisms in Behavior*, edited by L. Jeffress, Wiley, 1951.

[288] A. Ilachinski, *Cellular Automata: A Discrete Universe*, World Scientific Press, 2001.

[289] "Rule codes" are the the base-10 equivalent of a CA's binary f-rule specification. That is, if f is defined by (0,0,0) → a, (0,0,1) → b, (0,1,0) → c, (0,1,1) → d, (1,0,0) → e, (1,0,1) → f, (1,1,0) → g, (1,1,1) → h, then the rule code for f is given by $a\cdot 2^0 + b\cdot 2^1 + c\cdot 2^2 + d\cdot 2^3 + e\cdot 2^4 + f\cdot 2^5 + g\cdot 2^6 + h\cdot 2^7$. Hence, (0,0,0) → 0, (0,0,1) → 1, (0,1,0) → 1, (0,1,1) → 1, (1,0,0) → 0, (1,0,1) → 1, (1,1,0) → 1, (1,1,1) → 0 yields rule code = $0\cdot 2^0 + 1\cdot 2^1 + 1\cdot 2^2 + 1\cdot 2^3 + 0\cdot 2^4 + 1\cdot 2^5 + 1\cdot 2^6 + 0\cdot 2^7 = 110$.

[290] S. Wolfram, *A New Kind of Science*, Wolfram Media, 2002.

global behaviors of this (seemingly "trivial") rule cannot be predicted![291] For example, a well-known theorem from computer science, the *Halting Theorem*, asserts that there cannot exist a general algorithm for predicting when a computer will halt its execution of a given program.[292] Given that rule-110 is a universal computer—so that the *Halting Theorem* applies—one cannot, in general, predict whether a particular starting configuration of *on* and *off* sites will eventually die out. No shortcut is possible, even in principle. Rule-110, like all computationally universal systems, *effectively defines the most efficient simulation of its own behavior.*

More generally, if one is interested in understanding *why* something happens in a CAS, it is rarely true that the reason turns out to be a "simple" linear chain of causal events of the form: event A → "causes" event B → "causes" event C and so on. The "reason" why any event X occurs in a CAS is more typically: (1) not a priori "obvious" or easily discernable by direct observation alone (requiring focused study and/or experimentation), and (2) due to a *causal network* of events that are spatially and temporally coupled with X. Moreover, the parts and links in this causal network are themselves "meaningful" (locally, to X's existence, and, globally, to the entire system) only within certain dynamic contexts. Parts and relationships in CASs are seldom fixed throughout the time-evolution of CAS, and the nature of interrelationships can change depending on a given dynamic context and overall state of an evolving system.

Being able to "explain" (the reasons for) a certain behavior or phenomenon in a CAS—perhaps by generating, or "growing," it by using an MBM—does not necessarily imply that one is able to predict future behaviors. By the same token, while a CAS's behavior may be unpredictable—even in principle, a property attributable to, and reflective of, an inherent "irreducible complexity" typical of complex systems—this does not preclude the possibility to understand, or explain critical behaviors of, that same system. For example, while evolutionary theory does not "predict" the phenotypes that are observed in nature, it is perfectly adequate to "explain" the phenomenon of species diversity. Complexity thus suggests, as a general "lesson" to modelers, that less attention should be given to predicting specific behaviors, and more to understanding the underlying causal mechanisms behind emergent classes

[291] This is true strictly only for infinite systems, or, more precisely, for finite systems for which the area to which nonzero sites may be assigned can adaptively grow in unbounded fashion. However, the essence of our argument remains unchanged, since we are using rule-110's universality only to emphasize the contrast between the (seemingly) trivial specification of a local behavioral ruleset and what is arguably the ***most*** *complex imaginable behavior*—namely that of computational universality—that emerges on a global level.

[292] C. Moore and S. Mertens, *The Nature of Computation*, Oxford University Press, 2011.

of behaviors. This has potentially significant ramifications for the testing and evaluation of CAS-based algorithms underlying autonomous systems.

Since many of the most important behaviors of CAS appear when a system is (sometimes, very far) *out of equilibrium*, knowing (or "solving for") equilibrium states is often far less important than understanding the range of possible fluctuations and the nature and frequency of extreme events. For example, much of what is currently known about the statistics of traffic jams in vehicular dynamics comes not from solving differential equations describing continuum fluid flow, but from high-fidelity MBMs of traffic flow.[293] Likewise, it is unlikely that there will be an all-encompassing "theory" of autonomous systems, or set of equations that will predict the possible range of robotic-swarm behaviors. To the extent that *any* specific behaviors can be predicted about robotic swarms prior to deployment, MBMs will almost certainly play a key role.

Self organized criticality

Some lessons from the study of CAS are more subtle, and require one to delve a bit deeper into the literature on complex systems theory. One example is that of *self-organized criticality* (SOC),[294] which embodies the idea that dynamical systems with many degrees of freedom naturally self-organize into a critical state in which *the same events that brought that critical state into being can occur in all sizes, with the sizes being distributed according to a power-law*. SOC seeks to describe the underlying mechanisms for structures in systems that look like equilibrium systems near critical points but are not near equilibrium. Instead, they continue interacting with their environment, "tuning themselves" to a point at which critical-like behavior appears. In contrast, thermodynamic phase transitions usually take place under conditions of thermal equilibrium, where an external control parameter such as temperature is used to tune the system. Introduced in 1987 by Bak, Chen, and Wiesenfeld,[295] SOC is arguably the only existing holistic mathematical theory of self-organization in CAS, describing the behavior of many real systems in physics, biology, and economics.

[293] K. Nagel and S. Rasmussen, *Traffic on the Edge of Chaos*, Working Paper 94-06-032, Santa Fe Institute, 1994.

[294] H Jensen, Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems, Cambridge University Press, 1998.

[295] P. Bak, C.Tang, and K.Wiesenfeld, "Self-organized criticality: an explanation of 1/f noise," *Physical Review Letters* 59, 1987.

"Self-organized" refers to the fact that the system, after some initial transient period, naturally evolves into a critical steady state without any tuning of external parameters. This stands in marked contrast with the critical points at phase transitions in thermodynamic systems that can only be reached by a variation of some parameter (temperature, for example). "Criticality" refers to a concept borrowed from thermodynamics. Thermodynamic systems generally get more ordered as the temperature is lowered, with more and more structure emerging as cohesion wins over thermal motion. Thermodynamic systems can exist in a variety of phases—gas, liquid, solid, crystal, plasma, etc.—and are said to be *critical* if poised at a phase transition. Many phase transitions have a critical point associated with them, that separates one or more phases. As a thermodynamic system approaches a critical point, large structural fluctuations appear despite the fact that the system is driven only by local interactions. The disappearance of a characteristic length scale in a system at its critical point, induced by these structural fluctuations, is a characteristic feature of thermodynamic critical phenomena and is universal in the sense that it is independent of the details of the system's dynamics. Other than the absence of any control parameters, the resulting behavior is thus strongly reminiscent of the critical point in thermodynamic systems undergoing a second-order phase transition.

In general, SOC appears in systems that have the following properties: (1) many degrees of freedom; (2) strong local interactions; (3) number of parts is usually conserved; (4) slowly driven by exogenous "energy" source; and (5) energy is rapidly dissipated within the system. In systems that have these properties, SOC itself is characterized by: (1) a self-organized drive towards the critical state; (2) intermittently triggered (avalanche-style) release of energy; (3) sensitivity to initial conditions[296] (i.e., the trigger can be very small); and (4) the critical state is maintained without any external "tuning." Notably, SOC systems characteristically display *fat-tailed behavior*—i.e., "rare events" occur much more frequently than what is expected from normally distributed events in non-SOC dynamical systems. While the relevance of SOC for the design of (and understanding of the possible behaviors

[296] Note that sensitivity to initial conditions is usually a trademark of chaos in dynamical systems. Unlike fully chaotic systems, however, in which nearby trajectories diverge exponentially, the distance between two trajectories in systems undergoing SOC grows at a much slower (i.e. power-law) rate. Systems undergoing SOC are therefore only weakly chaotic. There is an important difference between fully developed chaos and weak chaos. Fully developed chaotic systems have a characteristic time scale beyond which it is impossible to make predictions about their behavior. However, no such time scale exists for weakly chaotic systems, so that long-time predictions may still be possible.

of) robotic swarms is, as yet, unclear, it is has been shown to be a fundamental driver of the dynamics of combat.[297]

Understanding a complex system's behavior requires both *analysis* and *synthesis*. The traditional Western scientific method is predicated on a fundamentally reductionist philosophy that assumes that the properties of a system may be deduced by decomposing the system into progressively smaller and smaller pieces. However, by analyzing a system in this way, there is strong chance that the most interesting emergent properties of the system will become lost; the chance of this happening only increases as the complexity of a system's behavior increases. Think of the absurdity of searching for consciousness by stripping a brain down to a few neurons! In meticulously probing the parts, the analytical-reductionist method inevitably loses sight of the whole. The understanding of complex systems also requires that a complementary holistic, or constructionist, approach be undertaken—in parallel with reducing a system down to its essential parts—in which one explores how the system's parts synthesize the whole.

Part of the power, and allure, of using MBMs to study complex systems (discussed in the next section), is that they embody precisely the kind of generative tools that a synthetic analysis of a complex system requires. They are designed to help build and understand complex systems, from the bottom up, and allow analysts to explore their (typically vast multidimensional) behavior space.

The traditional search for "optimal solutions" (to equations that describe a complex system's dynamics) is typically untenable. Emergent CAS behaviors are often due to a relatively small set of primitive rules and behaviors that constrain and define the dynamics of the system on its lowest levels. It is in this sense that a fundamental understanding of the critical nonlinear feedbacks that drive a system is essential to understanding how the whole CAS behaves. However, it is often the case that the components of a CAS need not be described in great detail in order for a model of the system to yield certain aggregate behaviors of interest; which has obvious ramifications for the design of robotic swarms, and is discussed in a later section. Important insights into system behavior can often be gained even in the absence of detailed knowledge of a given system, so long as the critical drivers and nonlinear relationships have all been properly identified.

Arguably, the most important qualitative lesson that complexity theory offers the prospective modeler (and/or autonomous system/swarm engineer / human operator)

[297] D. Roberts and D. Turcotte, "Fractality and Self-Organized Criticality of Wars," *Fractals: Complex Geometry, Patterns, and Scaling in Nature and Society* 6, no. 4, 1998; M. Lauren, "Describing Rates of Interaction between Multiple Autonomous Entities: An Example Using Combat Modelling," https://arxiv.org/abs/nlin/0109024.

is that the most useful CAS models are those that are developed specifically to support *exploratory modeling*. Exploratory modeling[298] refers to the practice of using models or simulations not as predictive vehicles, per se (in which inputs are used merely as "seeds" for generating predictions of specific outcomes), but rather as interactive tools for aiding analysts in conducting computational experiments and exploring (what might be loosely called) "plausibly possible futurescapes" of system behavior. CAS simulations are rarely developed solely to make predictions; rather, they more typically provide a conceptual underpinning to allow the analyst to interactively generate explanations of systemic behavior. Epstein[299] lists no fewer than sixteen reasons other than prediction for using models of CAS (the first of which is to help "explain" observed behavior).

The "best" CAS-based models are those that *instantiate* (in source code and through a model's visualization algorithms) an analyst's own (possibly abstract) theories and assumptions about a system and a system's behavior. The model's main purpose, in this case, is to do nothing other than what the analyst herself might have done—had the analyst been equipped with a massive store of "memory" and very fast "processor"—*only much faster*, and for a far greater number of scenarios and starting conditions. Exploratory modeling is especially useful for gaining insights into systems about which there may be significant uncertainties (such as is often the case in studies of CASs, in which there are myriad uncertainties regarding both low-level dynamics and high-level emergent behaviors). In effect, exploratory modeling may be viewed as an interactive search for plausible possible futures over a *space of models*, wherein each model represents a plausible distillation of a given system, given whatever a priori knowledge or past experience an analyst brings to bear on a particular problem.[300] In sharp contrast to traditional forms of modeling (wherein the focus is to use methods such as sensitivity analysis to estimate the variance of predictions of a model to its inputs), exploratory modeling—especially when aided by

[298] S. Bankes, "Exploratory modeling for policy analysis," *Operations Research* 41, no. 3, May/June 1993.

[299] J. Epstein, "Why Model?" *Journal of Artificial Societies and Social Simulation* 11, no. 412, 2008. The 16 reasons are: (1) explain (very distinct from predict); (2) guide data collection ; (3) illuminate core dynamics; (4) suggest dynamical analogies; (5) discover new questions; (6) promote a scientific habit of mind; (7) bound outcomes to plausible ranges; (8) illuminate core uncertainties; (9) offer crisis options in near-real time; (10) demonstrate tradeoffs / suggest efficiencies; (11) challenge the robustness of prevailing theory through perturbations; (12) expose prevailing wisdom as incompatible with available data; (13) train practitioners; (14) discipline the policy dialogue; (15) educate the general public; and (16) reveal the apparently simple (complex) to be complex (simple).

[300] S. Bankes and J. Gillogly, *Exploratory Modeling: Search Through Spaces of Computational Experiments*, RAND, RP-345, 1994.

multiagent-based simulations—seeks only to find plausible future states and trajectories of a system that are consistent with what is known.[301]

Multiagent-based models

Multiagent-based models (MBMs) represent a broad class of modeling techniques that have been developed specifically for studying complex adaptive systems[302] and, more recently, social dynamic systems.[303] MBMs are predicated on the basic idea that the ostensibly complicated global behavior of a CAS derives, collectively, from simpler, low-level interactions among the parts (or *agents*); indeed, MBMs are, by design, *explicit distillations of whatever particular complex system is being studied*. Essentially everything that is currently known about the general behavior of complex systems, including the properties of specific systems, has been derived from MBMs.

Roughly speaking, any real physical system that exhibits the following four basic properties is, in principle, amenable to being "modeled" by an MBM:

- *Heterogeneity*, meaning that the system consists not just of "exact copies of the same" part, but a variety of dissimilar, related parts with a mix of properties
- *Autonomy*, so that the behavior of the system, as a whole, is not dictated top-down by one central "controlling" agent, but rather is a collective property of the combined decisions of otherwise autonomous intelligent agents
- *Bounded rationality*, which prevents any part of the system from "knowing" what the entire is system is "doing" at any one time (so that each agent has access only to limited information and a finite store of computational resources by which to make "decisions")
- *Local interactions*, which constrains agents' actions to local interactions with other neighboring agents (of course, this constraint can be lifted for cases where agents making up a robotic-swarm, say, communicate with one another)

[301] S. Bankes and J. Gillogly, *Validation of exploratory modeling*, RAND, RP-298, 1994.

[302] G. Weiss, editor, *Multiagent Systems*, MIT Press, 2000.

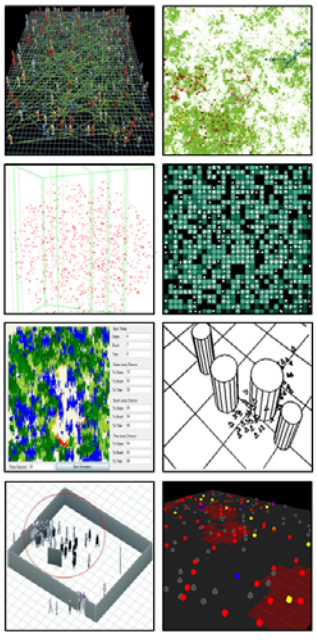
[303] J. Epstein, *Generative Social Science*, Princeton University Press, 2007; C. Nikolai and G. Madey, "Tools of the Trade: A Survey of Various Agent Based Modeling Platforms," *Jour. of Artificial Societies and Social Simulation* 12, no. 2, 2009.

A fifth property—*adaptive evolution*, by which agents are able to learn and grow as a simulation unfolds—is often part of an MBM; though, strictly speaking, a system does not need to be a "learning" system in order to be modeled as an CAS.

MBMs are rarely developed purely as predictive models—that is, as tools to produce an "expected outcome" (or behavior) of a system given some input. Instead, they are best used to provide *generative explanations*. Insights about the real system are gained by examining the emergent structures within the simulation. Traditional mathematical and computer models focus on high-level descriptions because they assume that complex behaviors observed on the aggregate level of a system must be described by "complex" rules. In contrast, MBMs proceed from the assumption that complex global behavior is often a self-organized emergent consequence of the intertwined web of much simpler, low-level interactions among (also, typically, "simple") individual agents. Of course, neither traditional models nor MBMs account equally well for all kinds of dynamical systems; however, if the system of interest is a CAS, MBMs are, by far, the most suitable modeling paradigm. MBMs are particularly well suited for simulating the self-organized emergent dynamics of complex networks.

Figure 16 shows a partial list of some landmark / prototype MBMS that have been developed across multiple problem domains.

Figure 16. A partial list of some landmark / prototype MBMs



Landmark / prototype agent-based models

- Boids (Bird flocking, Reynolds, 1987)
- SimCity (Urban dynamics, Maxis, 1989)
- Tierra (Evolution, Ray, 1992)
- ECHO (Ecologies, Holland, 1996)
- TransSim (Traffic flow, Barrett, 1997)
- Sugarscape (Epstein, Axtell, 1996)
- ISAAC (Land combat, CNA, 1996)
- BacSim (Bacterial growth, Kreft, et al., 1998)
- Legion (Crowd dynamics, Still, 2000)
- EMCAS (Electric power markets, 2002)
- EpiSims (Epidemics, Los Alamos, 2004)
- SOTCAC (Terrorist networks, CNA, 2006)

Meta-MBM toolkits

- Ascape (Brookings Institute)
- MASON (George Mason Univ)
- Mathematica (Wolfram Research)
- Netlogo (Northwestern University)
- Repast (Univ. of Chicago / Argonne National Lab)
- StarLogo (Media Laboratory, MIT)
- Swarm (Santa Fe Institute / Univ. of Michigan)

Ref: U. Wilensky and W. Rand, *An Introduction to Agent-Based Modeling*, MIT Press, 2015

"Generative explanations" proceed from the bottom up. By effectively decoupling individual rationality from macroscopic equilibrium, MBMs represent a new hybrid theoretical-computational tool[304]—one that is neither totally deductive nor totally inductive. While MBMs, like deductive methods, start from a set of primitive rules and assumptions, they do not prove theorems; rather, they generate behaviors that must themselves be further studied inductively. Unlike traditional induction, however, via which patterns are uncovered by studying empirically derived data, MBMs provide the framework for discovering (presumably real-world) high-level patterns that emerge out of the aggregate behaviors of lower-level entities and interaction rules. Traditional models ask, effectively, "How can I characterize the system's top-level behavior with a few, equally top-level, variables?" MBMs ask, instead, "What low-level rules and what kinds of heterogeneous, autonomous agents do I need to have in order to synthesize the system's observed high-level behavior?"

While a valid and useful answer to the first question can often be found, there is at least one significant drawback to this approach: so many simplifying assumptions must usually be made about the real system in order to render the top-level problem a soluble one, that other natural, follow-up questions, such as *"Why do specific behaviors arise?"* or *"How would the behavior change if the system were defined a bit differently?"* cannot be meaningfully addressed without first altering the set of assumptions. An analytical, closed-form "solution" may describe a behavior; however, it does not necessarily provide an explanation for that behavior. Indeed, subsequent questions about the behavior of the system must usually be treated as separate problems.

As a general class of problem-solving tools, MBMs span a wide spectrum of function and utility. Ranked roughly according to conceptual depth and sophistication, they can be used as:[305]

- *Classical simulation tools:* they can be used as dynamic representations, and therefore effectively provide instantiated "checks," of otherwise "conventional" closed-form solutions to tractable problems.

[304] J. Epstein, "Agent-Based Computational Models and Generative Social Science," *Complexity* 4, no. 5, 1999; and R. Axtell, *Why agents? On the varied motivations for agent computing in the social sciences,* Brookings Institution, Working Paper 17, November 2000.

[305] For additional details see J. Epstein, "Agent-Based Computational Models and Generative Social Science," *Complexity* 4, no. 5, 1999; R. Axtell, *Why Agents?* Brookings Institution, 2000; and J. Epstein, "Why Model?" *Journal of Artificial Societies and Social Simulation* 11, no. 4, 31 October 2008.

- *Logical inference engines:* they help identify—and discover—the logical consequences of the assumptions and/or constraints imposed on a problem or class of problems.

- *Sensitivity analysis:* they may be used to take systematic excursions away from equilibrium solutions derived by other means, thereby helping to characterize the solution space.[306]

- *Interactive theorizers:* MBMs may be used for helping to explain "why" something happens, and to enhance understanding of existing problem spaces.

- *Generators of "plausibly possible" futurescapes:* MBMs are generally best used as tools for exploring (the typically vast, multidimensional space of) "plausibly possible" future states of a system; i.e., as conceptual / analytical / visual aids for exploratory modeling. "Exploratory modeling" refers to the practice of using models or simulations not as predictive vehicles, per se (in which inputs are used merely as "seeds" for generating predictions of specific outcomes), but rather as interactive tools for aiding analysts in conducting computational experiments and exploring all possible future states of a system. They may also be used to gain insight into the relationship between low-level rules and behaviors and high-level emergent phenomena.

- *Experimental probes:* MBMs may be used to discover alternative and/or novel measures of effectiveness or measures of performance (of, say, specific adaptation or mitigation strategies), and to act as conceptual testbeds for suggesting real-world experiments.

- *Meta-modeling environments:* there are a growing number of meta-modeling platforms that, by deliberately minimizing the "details" about specific systems or problem domains, facilitate the design and development of MBMs. Some well-known examples include Swarm, MASON, Repast, StarLogo, and NetLogo. *OpenABM*—a consortium of researchers, educators, and professionals interested in MBMs (founded in 2007)—maintains an archive of meta-modeling platforms.[307]

[306] For example, CNA's EINSTein MBM simulation of combat was conceived for precisely this reason; namely, as a vehicle to take systematic excursions away from behavior described as closed-form solutions to the Lanchester equations. Ref: J. Taylor, *Force-on-Force Attrition Modelling*, Operations Research Society of America, Military Applications Section, 1980.

[307] https://www.openabm.org/modeling-platforms.

MBMs are particularly powerful modeling and simulation tools to use on systems that include a human decision-making and/or social/cultural component. Particularly important in the context of this report is the fact that MBMs are equally as adept at providing insights into what a system does, collectively, in terms of what its constituent parts do, as it is in providing insights into what happens to the individual components of a system because of what the systems does.[308]

Words of caution

While MBMs are inarguably powerful tools for understanding CASs (an assertion that is underscored, if not proven, by the breadth and depth of their application in a wide variety of subject-matter domains), their utility is not without costs. For example, despite many attempts at codifying fundamental principles (e.g., by Epstein,[309] Holland,[310] and Weiss[311]), CAS theory—the universally agreed upon conceptual framework that underlies all MBMs—itself remains more of a "work in progress" than final all-encompassing theory.[312] As of this writing, no universal definitions of "complexity" and "emergence" exist (indeed, the meaning of these terms are still hotly debated);[313] and there remains disagreements on core concepts and methodologies.

Most significantly, in the context of this paper, is that Verification, Validation, and Accreditation (VV&A) for MBMs is less than satisfactory, at best (as an objective set of standard practices), and even less universally agreed upon than CAS theory, at worst. In practice, each MBM developer effectively applies her own home-grown version of whatever VV&A technique seems appropriate for a specific type of model; and the veracity of each approach is subject to vagaries of chance acceptance and peer review

[308] S. Railsback and V. Grimm, *Agent-Based and Individual-Based Modeling: A Practical Introduction*, Princeton University Press, 2011.

[309] J. Epstein, *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science*, Princeton University Press, 2013.

[310] J. Holland, *Signals and Boundaries: Building Blocks for Complex Adaptive Systems*, MIT Press, 2014

[311] G. Weiss, *Multiagent Systems*, Second Edition, MIT Press, 2013.

[312] N. Johnson, *Simply Complexity: A Clear Guide to Complexity Theory*, Oneworld, 2009.

[313] M. Bedau and P. Humphreys, editors, *Emergence: Contemporary Readings in Philosophy and Science*, MIT Press, 2008; P. Humphreys, *Emergence*, Oxford University Press, 2016.

by subject matter experts (at least for cases in which whatever research the MBM was developed in support of is published in an accredited journal).[314]

One noteworthy VV&A-like technique that has received attention in the MBM community is that of *aligning* (or "docking") two or models. First proposed by Axtel at al.,[315] and Axelrod,[316] in the late 1990s, the idea is to see whether two models can produce the same results; i.e., using one model to "check" the output of another. The authors illustrate this concept by using, as their test-bed simulations, a model of cultural transmission designed by Axelrod[317] and the *Sugarscape* MBM of evolution that takes place on a notional "sugar" field, developed by Epstein and Axtell.[318] Since the models differ in many ways (and have been designed with quite different goals in mind), the comparison was not an especially easy one to make. Nonetheless, the authors report that the benefits of the sometimes arduous process of "alignment" far outweighed the hardships. In the end, the user communities of both models benefited from the alignment process by gaining a deeper understanding of how each program works, of their similarities and differences, and of how the inputs and outputs of each program must be modified before a fair comparison of what "really happens in either model" can be made.

UASs as CASs

On a fundamental level, UASs are prototypical complex adaptive systems, consisting of multiple nonlinearly interacting components such that the aggregate behavior is an emergent function of the entwined adaptive dynamics of their individual parts.[319]

[314] For example: *The Journal of Artificial Societies and Social Simulation*: http://jasss.soc.surrey.ac.uk/JASSS.html; *Advances in Complex Systems*: http://www.worldscientific.com/page/acs/aims-scope; and *Swarm Intelligence* (Springer-Verlag): http://www.springer.com/computer/ai/journal/11721.

[315] R. Axtel, et al., "Aligning simulation models: a case study and results," *Computational and Mathematical Organization Theory* 1, 1996.

[316] R. Axelrod, *The complexity of Cooperation: Agent-Based Models of Competition and Cooperation*, Princeton University Press, 1997.

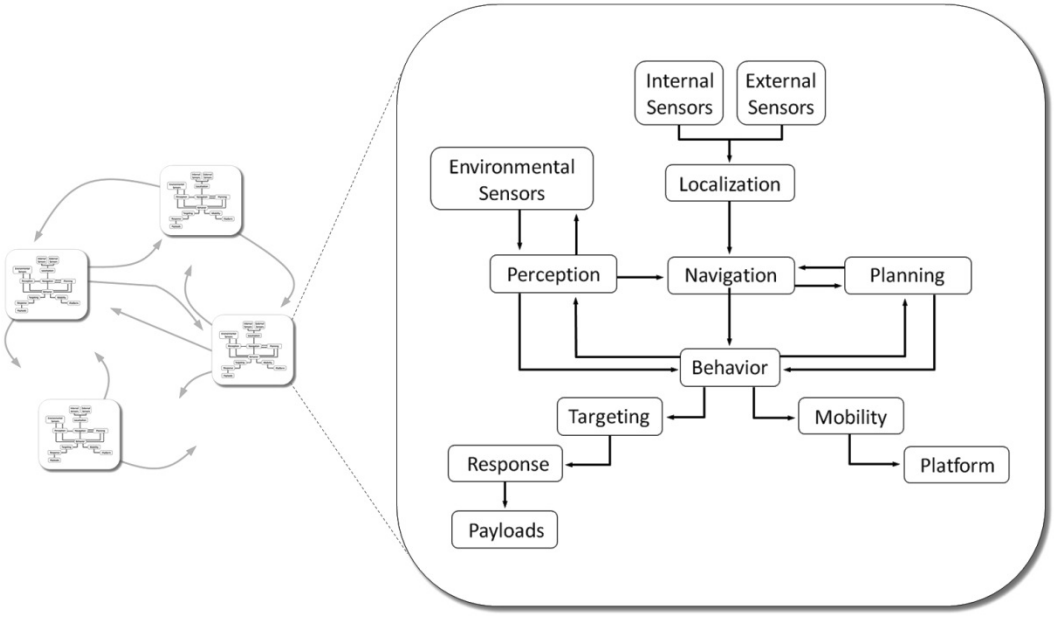
[317] R. Axelrod, "The dissemination of culture: a model with global polarization," *Journal of Conflict Resolution* 41, 1997.

[318] J. Epstein and R. Axtell, *Growing Artificial Societies*, MIT Press, 1996.

[319] F. Macias, "The test and evaluation of unmanned and autonomous systems," *ITEA Journal* 29, 2008.

Figure 17 shows, schematically, some of the key functional components and relationships of various subsystems that make up a typical UAS (excluding the part that involves human interaction and communications, but which will be described shortly). The left-hand-side of the figure highlights, notionally, the fact that an individual UAS may also be part of swarm, which entails an additional layer of "complexity." In this section we will only briefly touch upon this aspect of UASs, leaving a detailed discussion to a later chapter (see **Robotic Swarms**).

Figure 17. Key functional components and relationships of an autonomous unmanned system (*excluding* human interaction and communications)



After figure 3.1 in A. Finn and S. Scheding, *Developments and Challenges for Autonomous Unmanned Vehicles*, Springer-Verlag, 2010.

The various parts of an UAS are both numerous and feature different capabilities and characteristics:[320]

- *Internal sensors*: used to measure, for example, wheel velocity (odometers), steering angle, ground (radar or laser) or sea floor (acoustic) Doppler or depth/altitude (pressure sensors). Such sensors are, generally speaking,

[320] Chapter 3 in A. Finn and S. Scheding, *Developments and Challenges for Autonomous Unmanned Vehicles*, Springer-Verlag, 2010.

proprioceptive; i.e., they perceive internal factors that are effected by the environment and the UAS's own behavior.

- *External sensors:* which provide data regarding the position and orientation of the UAS in some absolute frame of reference (e.g., inclinometer, magnetic compass, and GPS). Also usually considered to be proprioceptive.

- *Environmental sensors:* which provide data (e.g. via radar, LADAR, EO, IR and acoustic sensors) to the UAS's perception algorithms, allowing the UAS to observe and develop a map of its environment. Also called *exteroceptive*; i.e., they perceive external factors that are not under the control of the UAS.

- *Localization:* provides estimates of the UAS's position, velocity, attitude, altitude rate, and acceleration.

- *Perception:* which is responsible for capturing, representing, and interpreting relevant environmental cues (e.g. location, geometry, spectral content, etc.), as observed by sensors, and relating these to features in the real world for the vehicle's moment-to-moment control, mission and task planning, payload control, etc.

- *Navigation:* which is responsible for generating a map of the UAS's local environment, a path to navigate the vehicle from its current location to the next waypoint or final destination, and the detection of any hazards that might impeded the UAS's progress. This module processes data from the localization and perception functions, and uses this information in conjunction with the behavior function to execute its mission.

- *Planning:* defines the algorithm that generates the sequence of actions the UAS must take from a specified starting position to its final destination (or activity) while avoiding obstacles and other "unanticipated" impediments. While the planning function is not directly linked with the UAS's sensory input, it may rely on other a priori forms of data (e.g., maps and mission objectives).

- *Behavior:* translates the combined outputs of the navigation, planning, and perception functions into actuator commands that allows the UAS to execute specific actions (e.g., move and/or fire weapons).

- *Mobility:* refers to the ability of the vehicle to traverse its environment. For unmanned ground vehicles, it is often expressed in terms of the size of a physical object that the vehicle can negotiate. For air and sea-based vehicles, mobility typically includes the aero– and hydro–dynamic properties of the platform.

- *Targeting:* in the event that the UAS is weaponized, the targeting module is responsible for aiming the weapon. If the UAS capable of autonomously deploying its weapons (an issue that will examined in detail in subsequent sections), the targeting module would likely be tightly coupled with both the planning and behavior modules.

- *Response:* which is a generic label for the module responsible for the tasking of the UAS's payloads to react to and/or engage with objects within the UAS's environment (e.g., firing a weapon, or panning, tilting & zooming an ISR payload to create or enhance a situational awareness picture for the user, or to respond to and engage with objects within this picture

Additional components (both those indicated on the figure and not) include *payloads* (e.g., radar, electronic warfare equipment, mine countermeasures, and weapons); *energy* (the rate at which power is used is a key parameter); *propulsion systems* (which are typically designed around specific tasks and missions); and *health and usage monitoring systems* (HUMS), by which the UAS can self-monitor and (if necessary) diagnose issues that may lead to (or are a result of) a system malfunction.

Even a cursory comparison between the network of interacting parts and processes that make up a typical UAS and the basic "ingredients" that define a CAS suggests that UASs are prototypical complex systems: they are "agents" that respond and adapt to changing physical (and/or virtual) environments, act on only the local data or information that is available to them, and can alter their behavioral strategies based on the feedback they get either from their own sensors or from outside agents (including other UASs). If an individual UAS is also part of a swarm, the comparison becomes only that much more compelling.

On an even deeper level, since neither the function nor performance of UASs (nor of their key components) can be described *in situ*, but require broader contexts of human control and interaction, physical environment, and mission space to furnish meaning, UASs are also textbook examples of complex adaptive *System-of-Systems* (CASoS).[321] John Holland, a pioneer in the study of complex adaptive systems and an early member of the Board of Trustees and Science Board of the *Santa Fe Institute*,[322] likened a system-of-systems to an artificially created complex adaptive systems: "It is manufactured to achieve a predefined mission and will involve a large number of interacting entities with persistent movement and reconfiguration, changing based

[321] A. Fereidunian, et al., "A Complex Adaptive System of Systems Approach to Human-Automation Interaction in Smart Grids," in *Contemporary Issues in Systems Science and Engineering*, edited by M. Zhou, IEEE-Wiley Press, 2015.

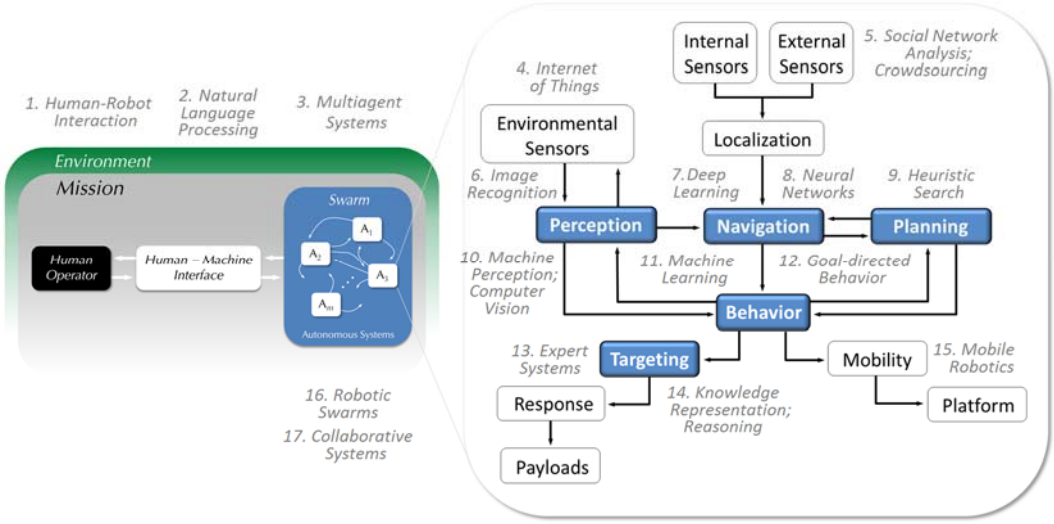
[322] https://www.santafe.edu/.

on changes in context, ordered through self-organization, with local governing rules for entities and increasing complexity as those rules become more sophisticated."[323]

Linking autonomy with AI

UASs are not only CASoS, but their design, function, and performance depends critically on the efficacy of various forms of embedded AI. Figure 18 overlays a (far from exhaustive) taxonomy of AI techniques and methods on the schematic block of a UAS's key functional components (shown in the previous figure), and adds explicit couplings to human interaction (i.e., human as operator and/or collaborator), physical environment, mission space, and other UASs.[324]

Figure 18. Generalizing a UAS as a CASoS and linking autonomy with AI



Right-hand-side of figure based on figure 3.1 in A. Finn and S. Scheding, *Developments and Challenges for Autonomous Unmanned Vehicles*, Springer-Verlag, 2010.

[323] J. Holland, *Hidden Order: How Adaptation Builds Complexity*, Addison-Wesley, 1995

[324] R. Brooks, *Cambrian Intelligence: The Early History of the New AI*, MIT Press, 1999.

Inherent "surprise" in complex systems

Charles Perrow, in his seminal book *Normal Accidents*,[325] argues that in tightly coupled complex systems, such as modern military weapon systems, accidents are "normal" events; that is, they are *inevitable*. The basic theme of Perrow's book—and what has come to be known as Normal Accident Theory (NAT)[326]—is that a priori innocuous individual and seemingly unrelated events accumulate and align to spawn major system malfunctions that can, in turn, induce catastrophic results. For example, in summarizing the forensic investigation of a CD-10 crash at Chicago O'Hare Airport in 1979, Perrow quotes from the National Transportation and Safety Board's (NTSB's) report:

> The loss of control of the aircraft was caused by the combination of three events: the retraction of the left wing's outboard leading edge slats; the loss of the slat disagreement warning system; and the loss of the stall warning system – all resulting from the separation of the engine pylon assembly. Each by itself would not have caused a qualified flight crew to lose control of the aircraft, but together during a critical portion of the flight, they created a situation, which afforded the flightcrew an inadequate opportunity to recognize and prevent the ensuing stall of the aircraft.

In his book, Perrow provides a wide set of examples to support his main thesis; e.g., air transport system, banking and financial systems, Three Mile Island (TMI), marine accidents, dams, hospitals, mines, petrochemical plants, weapon systems, and DNA research. Each of these systems is innately prone to failure because their function depends on the integrity of myriad interconnections among equipment, subsystems, operating procedures, human operators, and the environment. Despite any safeguards and redundancies that may be deliberately built into the system to act as buffers against possible malfunctions, it is inevitable (according to NAT) that there will be occasional failures, perhaps only minor—and all-but-invisible—ones unanticipated by the designers of a system (or via policies, procedures, and/or training) that collectively, via the network of systemic interactions, build one on the

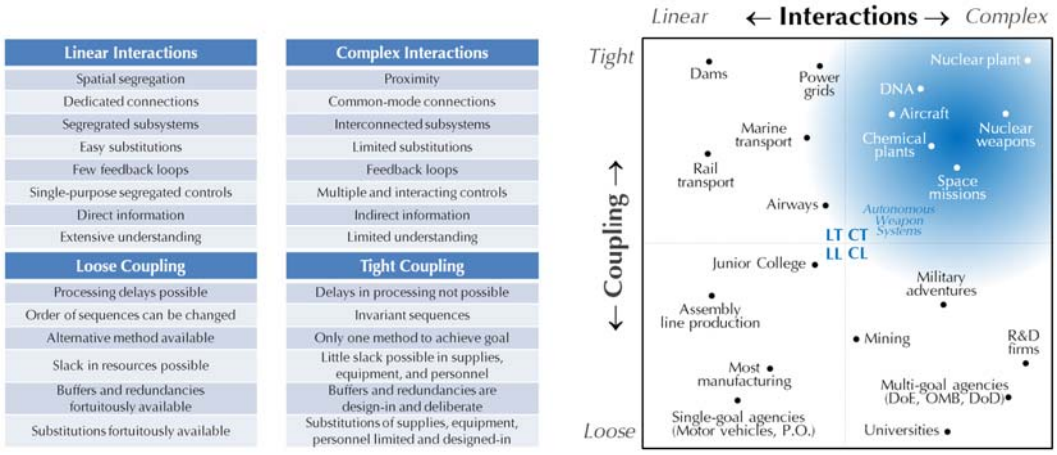
[325] Perrow was arguably the first to introduce the idea that accidents are inherent in the nature of complex systems, and that deliberate attempts (by humans in the loop) to avoid the consequences may actually engender rather than ameliorate catastrophic failure. Ref: C. Perrow, *Normal Accidents: Living with High-Risk Tech*nologies, Updated Edition, Princeton University Press, 1999.

[326] K. Weick, "Normal Accident Theory (NAT) as Frame, Link, and Provocation," *Organization & Environment* 17, no. 1, 2004.

other, causing failures to cascade, ultimately bringing part or all of the entire system down.

A system's propensity for (i.e., *risk* of) accidents depend on the *interactive complexity* of, and strength of *dynamic couplings* among, its individual elements. Systems with lots of parts that are linearly coupled (so that the magnitude of system-wide effects scale proportionately to the size of small perturbations within the system) are less prone to unanticipated behaviors than systems with the same (or even fewer) number of components that are nonlinearly coupled, and in which small local changes can induce disproportionately large global effects.

Figure 19. Characteristics of the two major variables in Perrow's Normal Accident Theory (NAT): *interactions* and *couplings*



Linear Interactions	Complex Interactions
Spatial segregation	Proximity
Dedicated connections	Common-mode connections
Segregrated subsystems	Interconnected subsystems
Easy substitutions	Limited substitutions
Few feedback loops	Feedback loops
Single-purpose segregated controls	Multiple and interacting controls
Direct information	Indirect information
Extensive understanding	Limited understanding
Loose Coupling	**Tight Coupling**
Processing delays possible	Delays in processing not possible
Order of sequences can be changed	Invariant sequences
Alternative method available	Only one method to achieve goal
Slack in resources possible	Little slack possible in supplies, equipment, and personnel
Buffers and redundancies fortuitously available	Buffers and redundancies are design-in and deliberate
Substitutions fortuitously available	Substitutions of supplies, equipment, personnel limited and designed-in

Figure 19 summarizes the characteristics of the two major variables, *interactions* and *coupling* (along the left-hand-side), and examples of where some systems generally fall within the two dimensional space (on the right). Interactions in a system span the space from *linear* to *complex* and coupling can be either *loose* or *tight*:[327]

- LL — *Linear Interaction, Loose Coupling:* the system harbors few complex interactions; components that fail can be gracefully isolated without disrupting system function; accidents can be ameliorated in either a top-down manner from a central authority or a decentralized, bottom-up manner.

[327] T. Gale, "Normal Accidents," Encyclopedia of Science, Technology, and Ethics, 2005.

- LT — *Linear Interaction, Tight Coupling:* tightly coupled systems are those in which: (1) processes happen very fast and cannot be turned off, (2) the failed parts cannot be isolated from other parts, and (3) there is effectively a single path that leads to a successful outcome. Systems with tightly coupled linear interactions are predictable, but improvised workarounds to local failures are not generally possible, and must be explicitly designed into the system. These systems have little slack (measured in terms of the time between when the local behavioral perturbations occur and when some internal/external control can be applied to the system in order to prevent a cascading failure), and delays typically disrupt their global behavior.

- CL — *Complex Interaction, Loose Coupling:* system contains many complex interactions, with many control parameters and unplanned behaviors. Since system is loosely coupled, there is some slack in its behavior, and local adaptive "corrections" to malfunctions (such as finding alternative dynamic pathways for certain functions) will not necessarily disrupt the whole system, and may improve it.

- **CT — *Complex Interaction, Tight Coupling:*** tight coupling of a complex network of interactions makes any failure potentially disruptive; local ameliorations can also potentially result in a system-wide failure (they compromise the functional integrity of vital parts of the system). Most problems in such systems are fundamentally unpredictable because of the combinatorial complexity of parts and connections. The only effective route to amelioration is a decentralized one, whereby those parts (and/or operators) that are closest to each malfunctioning/failing subsystem undertake a slow, careful analysis of a given failure to determine what went wrong and what can be done about it.

We have highlighted the *complex tightly coupled systems* in bold to emphasize that it is this class of complex systems—into which autonomous weapon systems inarguably fall (encompassing, notionally, the area that is highlighted in blue in figure 19)—for which surprises and/or accidents will inevitably occur over a long enough period of time. These types of systems are the most prone to display "surprising behaviors" and to "fail" (to perform according to design specifications) because of the inherent unknowability of the complete set of global behaviors that can arise from their nonlinearly coupled parts. Failures can result via interactions within the system itself (e.g., the logic that defines how a robotic swarm ought to behave), via human-in-the-loop control (which may have unanticipated effects on an autonomous system's behavior), and/or via a system's dynamic coupling environment (in which a system may encounter contexts that were not anticipated by its designers). When there is sufficient slack (as defined above bullet describing LT systems), accidents can be ameliorated to avoid catastrophe. However, when the components of a system are tightly coupled, local failures can cascade rapidly from

one subsystem to another, and such systems offer little slack to act as a dynamic buffer. Accidents in complex tightly coupled systems thus become inevitable, or even "normal." The potential for catastrophic system failure increases as the complexity of interactions and coupling strengths both increase; i.e., the closer one gets to the upper right corner in the plot on right-hand-side in figure 19.

Control & risk of autonomy

NAT provides a useful conceptual framework for discriminating among different kinds of systems, generally, and for characterizing why autonomous systems, in particular, are likely to be prone to displaying "surprising" behaviors. Implicit in this framework is that as the complexity of a system increases (in both of NAT's dimensions), the ability to control the behavior of a system decreases; and any degradation of control entails *risk*. The risk in deploying an autonomous system is that the system may not perform the required operational task(s). Of course, some of the ways in which an autonomous system might fail to perform can be anticipated in advance (e.g., simple programming bugs, changing environmental conditions, human error), and human operators can, accordingly, adjust the system's behavior to accommodate these malfunctions. Other performance failures may not be easily anticipatable. Indeed, as we have argued here and elsewhere throughout this paper, a fundamental property of all sufficiently tightly and nonlinearly coupled complex systems is that their global behavior space contains behaviors that cannot in principle be predicted from knowing the local rules by which its components interact (short of exhaustively probing the system for all possible behaviors, which is combinatorically impossible except for the simplest systems). Understanding the conditions under which an autonomous system might fail—or, equivalently, the conditions under which an autonomous system *cannot be controlled* by a human operator—is essential to assessing the risks involved in deploying such systems. And, given that our current general understanding of complex adaptive systems is insufficient to a priori account for all possible global system behaviors (such an understanding may be fundamentally impossible to achieve), new analytical methods may need to be developed for assessing risk in employing autonomous systems.

Of course, even "simple" complex systems—that is, "automated," but not *autonomous*, systems that do not exhibit emergent behaviors or adapt and/or learn from past experiences—can spawn unexpected behaviors. Such behaviors can arise because of bugs in programming, flaws in engineering design, the sheer number of otherwise "simple to understand" parts (that may overwhelm the human operator),

and/or unanticipated interactions with the environment.[328] Other modes of failure can arise in military systems, since they operate *in inherently adversarial environments*. For example,[329] incomplete information, an accelerated pace of interactions, unanticipated interactions between adversarial systems, hacking (both by conventional means and by exploiting predictable behaviors), and spoofing (i.e., sending false data).

[328] In 2007, eight F-22s experienced a catastrophic Y2K-like computer failure when their onboard computer systems all shut down as they crossed the international dateline (the impact of which had not been identified in during software testing). Ref: "This Week at War," *CNN*, February 24, 2007: http://transcripts.cnn.com/TRANSCRIPTS/0702/24/tww.01.html.

[329] P. Sharre, "Autonomous Weapons and Operational Risk," *Center for a New American Security*, Feb 2016.

Robotic swarms

Robotic Swarms (RSs) refers to a young but burgeoning interdisciplinary research and development field that studies the collective cooperative dynamics of a large number of decentralized distributed robots through the use of "simple" local rules.[330] It is directly inspired by how nature forms, and exploits, swarms; e.g., in which societies of insects can perform tasks that are beyond the capabilities of any individual member.[331] In the following discussion, we will sometimes make a distinction (as per convention in the extant literature) between multi-robot systems (that contain a relatively few robots) and a bona-fide "swarm" (that contains a large number of robots; say, > 10); when the distinction does not matter, either type of system will be referred to generically as a "robotic swarm."

RS lies at the cusp of several interrelated research domains that have emerged over the last 20 to 30 years, including *AI*, *artificial life*,[332] *complex adaptive systems*,[333] and *particle swarm optimization* (which studies computational methods for finding close-to-optimal solutions to combinatorial optimization problems; see below).[334] RS also borrows from (and relies heavily on) *multi-agent based modeling* techniques to first simulate and understand the behaviors that must ultimately be instantiated in hardware.[335] Since it is impossible to provide anything but a cursory look at any of these important fields, our discussion will focus on only those aspects that are directly applicable to the study; references for deeper dives are provided as needed.

[330] E. Şahin, "Swarm robotics: from sources of inspiration to domains of application," in *Swarm Robotics Workshop: State-of-the-Art Survey*, E. Şahin and W. Spears, editors, *Lecture Notes in Computer Science*, Springer-Verlag, no. 3342, 2005.

[331] E. Bonabeau, G. Theraulaz, and M. Dorigo, *Swarm Intelligence: From Natural to Artificial Systems*, Oxford University Press, 1999.

[332] M. Komosinski and A. Adamatzky, editors, *Artificial Life Models in Software*, Second Edition, Springer-Verlag, 2009.

[333] N. Boccara, *Modeling Complex Systems*, Second Edition, Springer-Verlag, 2010.

[334] A. Hassanien and E. Emary, *Swarm Intelligence*, CRC Press, 2015.

[335] U. Wilensky and W. Rand, *An Introduction to Agent-Based Modeling*, MIT Press, 2015.

One of the earliest suggestions of applying natural swarms to warfare was by Libicki in 1995:[336]

> Today, platforms rule the battlefield. In time, however, the large, the complex, and the few will have to yield to the small and the many. Systems composed of millions of sensors, emitters, microbots and mini projectiles, will, in concert, be able to detect, track, target, and land a weapon on any military object large enough to carry a human. The advantage of the small and the many will not occur overnight everywhere; tipping points will occur at different times in various arenas. They will be visible only in retrospect.

Two groundbreaking monographs on the possible military benefits to swarming (albeit focused on swarms of conventional troops and weapons, not robots) were published in 2000[337] and 2005 by the Rand Corporation.[338] The latter, in particular, traces evolution of conflict across history: from melee, to mass, maneuver, and, finally, swarms.

A key ingredient in swarm robotics is that of *self-organization*; i.e., the emergence of macro-level behavior from local nonlinear interactions among individual agents, and between system components and their environment. Self-organization results from the combination of four basic elements:[339] *positive feedback*, *negative feedback*, *randomness*, and *multiple interactions*.

Swarms—whether software-based or instantiated in hardware—are inextricably entwined with the study of complex adaptive systems (CASs) and multiagent-based modeling (as discussed in the preceding section). Indeed, one of the first general-purpose simulation platforms for the study of complex adaptive systems is called *Swarm*,[340] developed at the Santa Fe Institute during the 1990s.[341]

[336] M. Libicki, "Mesh and the net: speculations on armed conflicts in an age of free silicon," McNair Paper 28, *National Defense University*, 1995.

[337] J. Arquilla and D. Ronfeldt, *Swarming and the Future of Conflict*, Rand Corporation, 2000.

[338] S. Edwards, *Swarming and the Future of Warfare*, Ph.D. Thesis, Pardee Rand Graduate School, 2004: http://www.rand.org/content/dam/rand/pubs/rgs_dissertations/2005/RAND_RGSD189.pdf.

[339] A. Ilachinski, *Artificial War: Multiagent-Based Simulation of Combat*, World Scientific, 2004.

[340] http://www.swarm.org/wiki/Swarm_main_page.

[341] N. Minar, et al., "The Swarm Simulation System: A Toolkit for Building Multi-Agent Simulations," Santa Fe Institute (SFI), SFI Working Paper 96-06-042, June 1996.

Examples of swarming in nature include *colonies of bacteria* (e.g., resistance to antibacterial agents by colonies of bacteria in biofilms has been observed to be 500 times more than that of individual bacteria of the same kind),[342] *schools of fish* (which fosters both foraging and help defend against predators),[343] *locust swarms* (for which it has been shown that very high densities of individuals in a swarm induce a phase transition disordered to highly aligned collective movement),[344] *bee colonies* (e.g., honeybee swarms allocate tasks dynamically and adaptively in response to changes in the environment),[345] *bird flocks* (e.g., the evolution of cooperation for foraging and migration groups)[346] and *bird swarms* (the "simple" modeling of which was pioneered by Craig Reynolds in the 1980s; see discussion below),[347] *termite colonies* (e.g., the massive mounds built by the termite *Macrotermes natalensis* are built in a decentralized manner by the cooperative "engineering" of thousands of individuals, yet regulate the environment of the entire colony as though they were colossal heart-lung machines),[348] and *ant colonies* (in which individual members cooperate by communicate using pheromones to mark paths from their nest to food sources).[349] Natural swarms range in size from a few individuals to highly organized colonies that occupy large spaces and consist of many millions of individuals. Specific group behaviors include path planning, nest construction, architectural engineering, and task allocation. Social insects exchange information (e.g., to communicate the location of a food source or the presence of a predator) *locally*; i.e., without either having or relying on any information about the global environment. The implicit communication through any dynamic changes made in the environment (such as by leaving a pheromone trail) is called *stigmergy*.[350]

[342] J. Costerton, et al., "Microbial biofilms," *Annual Review of Microbiology* 49, 1995.

[343] L. Fuiman and A. Magurran, "Development of predator defenses in fishes," *Reviews in Fish Biology and Fisheries* 4, no. 2, 1994.

[344] J. Buhl et al., "From disorder to order in marching locusts," *Science* 312, 2006.

[345] M. Beekman, et al., "What makes a honeybee scout?" *Behavioral Ecology and Sociobiology* 61, 2007.

[346] S. Shen, H. Reeve, and W. Herrnkind, "The brave leader game and the timing of altruism among nonkin," *The American Naturalist* 176, no. 2, 2010.

[347] C. Reynolds, "Flocks, herds, and schools: A distributed behavioral model," *Computer Graphics* 21, no. 4, 1987.

[348] J. Turner, *The Extended Organism: The Physiology of Animal-Built Structures*, Harvard University Press, 2002.

[349] D. Gordon, *Ant Encounters: Interaction Networks and Colony Behavior*, Princeton University Press, 2010.

[350] O. Holland and C. Melhuish, "Stigmergy, self-organization, and sorting in collective robotics," *Artificial Life* 5, no. 2, 1999.

Of course, on the largest scale, human culture as a whole may also be viewed as a massive swarm of swarms.[351] Human swarms (which can be of varying sizes and complexities) are facilitated, as are other forms in nature, by the cooperative sharing of information. Human swarms on smaller scales may spontaneously "self organize" when the conditions right. For example, in the 2011 London Riots,[352] real-time dissemination of locations of police barricades (via Blackberrys) allowed rioters to avoid authorities and adaptively re-organize in other areas to continue looting. However, perhaps somewhat counterintuitively, it is generally harder for humans to self-organize as a "swarm" than insects or animals. This is because one of the key ingredients in the dynamics of cooperative behavior is that a swarm's constituent agents need to be "simple" (in a way that will be defined more carefully below); i.e., human "agents" must be willing to relinquish "control" over anyone else's actions other their own. This is easy to do for "simple minded" robots (as we will see); but, generally, not so easy to achieve for groups of humans.

What do all natural swarms have in common? The answer is that each harbors certain key elements of a broader set of dynamical systems called complex adaptive systems (which are discussed in more detail later). The individual robots in a swarm are typically:[353]

- *Autonomous* (i.e., not under a centralized control)
- *Situated* in the environment (and can act to modify it)
- Capable of *sensing* their local environment and other nearby robots
- Able to *communicate* (locally) with other robots
- *Unaware of the global state* of the environment (and other robots)
- Able to *cooperate* with other robot to perform a given task(s)

In addition, and taking a cue from natural swarms, robotic swarms are designed to be *robust* (i.e., to retain the ability to perform their assigned task(s) after the loss of a few or more individuals), *scalable* (i.e., so that their ability to perform well is not strongly affected by group size), and *flexible* (i.e., so that they can cope with a broad spectrum of different environments and tasks).[354]

[351] W. Buckley, *Society: A Complex Adaptive System*, Routledge, 2013.

[352] "Getting to the root of the U.K. riots," CBC News, 9 Aug 2011: http://www.cbc.ca/news/world/getting-to-the-root-of-the-u-k-riots-1.1112695.

[353] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: A review from the swarm engineering perspective," *Swarm Intelligence* 7, no. 1, 2013.

[354] S. Camazine, et al., *Self organization in biological systems*, Princeton University Press, 2001.

Real robot swarms may differ in some ways from their natural-born cousins. One obvious difference is that where natural swarms evolve on their own, robotic swarms must be deliberately designed. We will later examine the methods and challenges inherent in the design process. Another difference is that while natural swarms tend to be homogenous, robotic swarms can be heterogeneous (i.e., involve a mix of different types of robots performing various tasks). Real swarms are also prone to "hacking" (and other cyber concerns), such as a recent incident in which the control of a commercial drone was hijacked via intrusion into an unencrypted Wi-Fi.[355]

The research literature is replete with surveys on the large number of basic swarm behaviors:[356]

- Aggregation / rendezvous
- Area coverage
- Flocking and formation
- Collective movement / transport
- Collective mapping
- Directed flocking
- Dispersion
- Pattern formation
- Navigation
- Simulating ant colonies
- Task allocation
- Obstacle avoidance
- Foraging / search algorithms

Arkin[357] provides a list of general advantages and disadvantages of multi-robotic systems over those of single-robot systems. Advantages include:

- *Improved performance:* if tasks can be decomposed and performed in parallel, groups can achieve tasks more efficiently

[355] Samy Kamkar, "Skyjack," *samy.pl*, December 2, 2013: http://samy.pl/skyjack/.

[356] D. Floreano and C. Mattiussi, *Bio-Inspired Artifical Intelligence*, MIT Press, 2008; M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: A review from the swarm engineering perspective," *Swarm Intelligence* 7, no. 1, 2013.

[357] R. Arkin, *Behavior-Based Robotics*, MIT Press, 1998.

- *Task enablement:* just as in natural systems, groups of robots can perform certain tasks that are impossible to accomplish for single robots
- *Distributed sensing:* a group of robots effectively forms a "sensor grid" from which the collective information is potentially far wider than what is possible via the sensor range of a single robot
- *Distributed action:* multiple simultaneous cooperative actions can be performed in different places at the same time
- *Fault tolerance:* the failure of a single robot within a group does not necessarily imply that a given task cannot be accomplished

Among the disadvantages and/or challenges of multi-robot systems are:

- *Interference:* the actions of individual robots in a group (even otherwise "coordinated" ones) may mutually interfere (due to collisions, occlusions, loss of communications, etc.)
- *Uncertainty concerning other robots' intentions:* coordination requires that each robot "fully understands" what other robots are doing and what they expect other (nearby) robots to be doing; in the event of uncertainty (or less than threshold level "clarity") robots can compete instead of cooperate
- *Overall system cost:* while single-robot systems are, typically, more "complex" (to account for the array of behaviors and actions that the multi-robot system presumably is being engineered to provide), and therefore also costlier, it is not a given that a swarm of individually less complex robots will, as a group (that performs required tasks at least as well as the presumptively "more complex" single-robot system), cost less than the single robot; for military-grade autonomous robots and robotic swarms (discussed later) the total life-cycle cost must include training, maintenance, and human operators

Swarm intelligence

Swarm intelligence (SI) is a catch-all phrase that refers to a large (and still growing) class of bio-inspired computational algorithms based on the decentralized cooperative behaviors of swarms of social insects, flocks of birds, schools of fish, and even the processes of natural evolution as a whole. Introduced over a half century ago,[358] modern incarnations of SI include evolutionary programming (EP), genetic algorithms (GA), genetic programming (GP), differential evolution (DE),

[358] L. Fogel, A. Owens, and M. Walsh, *Artificial Intelligence through Simulated Evolution*, John Wiley, 1966.

evolution strategy (ES), ant-colony optimization (ACO), artificial bee colony (ABC), harmony search (HS), and particle swarm optimization (PSO).[359]

In nature, the search for beneficial adaptations to a continually changing environment (i.e., "natural evolution") is fostered by the cumulative evolutionary knowledge that each species possesses of its forebears. This knowledge is encoded in the chromosomes of each member of a species, and is passed on from one generation to the next by a mating process in which the chromosomes of "parents" produce "offspring" chromosomes. SI techniques mimic these natural processes by relying on heuristics directly inspired by their natural counterparts: reproduction, mutation, recombination, and selection. Of course, the details by which the analogy between nature and computer algorithm is drawn depend on the specific SI technique being employed.

For example, GAs mimic the genetic dynamics underlying natural evolution to search for optimal solutions of general combinatorial optimization problems.[360] They have been applied to the traveling salesman problem, VLSI circuit layout, gas pipeline control, the parametric design of aircraft, neural net architecture, models of international security, and strategy formulation.[361] The basic idea behind GAs is very simple. Given a "problem"—which can be as well-defined as maximizing a function over some specified interval or as seemingly ill-defined and open-ended as evolution itself, where there is no a-priori discernible or fixed function to either maximize or minimize—GAs provide a mechanism by which the solution space to that problem is searched for "good solutions." Possible solutions are encoded as chromosomes (or, sometimes, as sets of chromosomes), and the GA evolves one population of chromosomes into another according to their fitness by using some combination (and/ or variation) of the genetic operators of reproduction, crossover and mutation.

ACO works by mimicking the way ants create and mark trails from their nest (N) to food (F) sources and back again (the general process is outlined in figure 20).[362] Foragers sense the pheromone markers and follow the path to food discovered by other ants, reinforcing the trail for still other ants to follow in their wake by depositing more markers. The shortest path to the food thus effectively forms over time as it is continually strengthened by positive feedback via the collective action of

[359] Z. Michalewicz and D. Fogel, *How to Solve It: Modern Heuristics*, Springer-Verlag, 2005

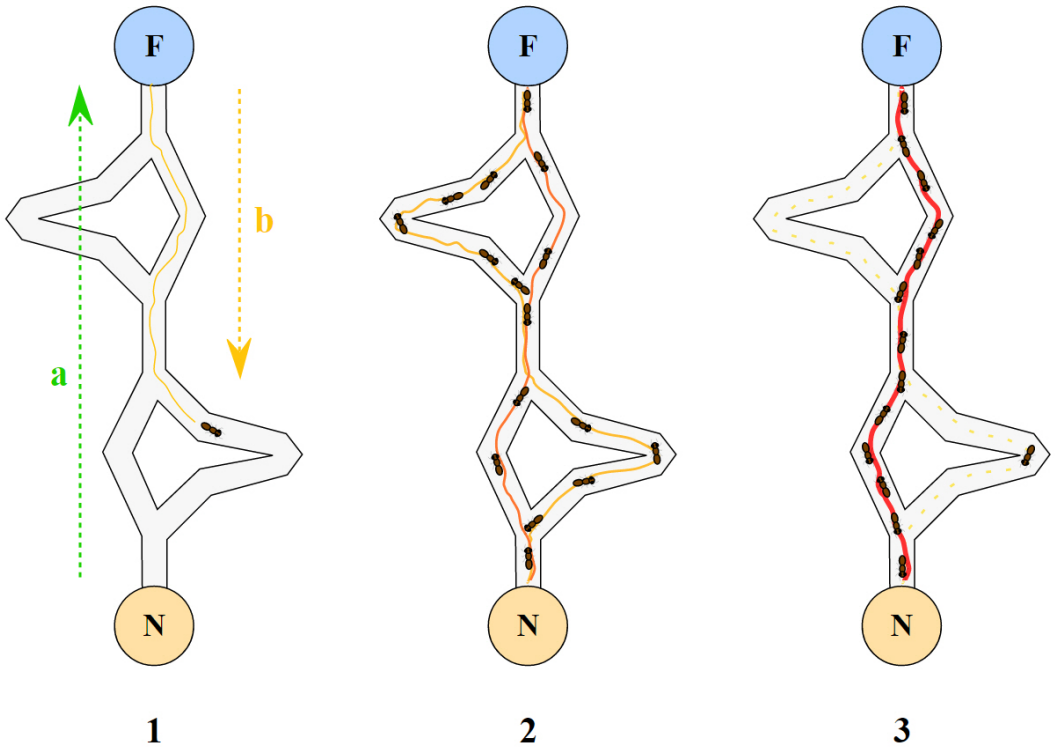
[360] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, 1998.

[361] A. Ilachinski, *Artificial War*, World Scientific Press, 2004.

[362] M. Dorigo, "Optimization, Learning, and Natural Algorithms," Ph.D. Thesis, Politecnico, di Milano, Italy, 1992; M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperating agents," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 26, no. 1, 1996.

the foraging ants. If the food source is exhausted and/or some obstacle along the path prevents ants from traveling toward the source, the extant path's strength immediately diminishes since no more pheromone is deposited on that path. In time, as the original pheromone evaporates, fewer and fewer ants take the old path; and, assuming some new path is discovered, more ants take the new path, thereby adaptively creating a new optimal path. The "takeaway" is that the "intelligence" (that leads to the discovery of any viable path) does not reside in any one individual ant, but is instead distributed among the entire group of foragers.

Figure 20. Schematic illustration of the Ant Colony Optimization (ACO) algorithm



Ref: J. Dreo, distributed under the provisions of the GNU Free Documentation License: https://upload.wikimedia.org/wikipedia/commons/a/af/Aco_branches.svg.

Another kind of swarm that occurs in nature, and is mimicked by ABC algorithms, is the honeybee swarm. Honeybees dynamically allocate tasks (e.g., tending the queen and brood, communicating, foraging, storing and distributing honey and pollen) and cooperatively adapt to changes in the environment. As foragers, individual bees take cues from the environment (odor, presence of pother bees, etc.) and rely on memory (e.g., recall of source of pollen and direction of odor). Similar to how ants leave pheromone deposits to attract other ants to a particular "good" path (i.e., one that leads to a food source), bees perform a "dance" on the area of the comb as a way of

communicating information about the food source (e.g., its richness, distance, and direction). The *type* of dance correspond to how far a given pollen source is: a *round dance* is performed if the distance is less than about 100 meters, a *waggle dance* if it is farther, and a *round dance* if there is no directional information. Longer distances cause faster dances. And a *tremble dance* is executed if a bee has determined that it will take a long time to deposit its nectar.[363] A honeybee colony also determines how many individual bees to assign to each of the many tasks that must be performed.

Some specific bee-inspired algorithms include: (1) the *Bees* algorithm,[364] which mimics the foraging behavior of bees (and consists, effectively, of combing local neighborhood search with a random global search), (2) the *BeeHive* algorithm,[365] which mimics how honeybees communicate (and is intended mainly as a routing algorithm), and (3) the *Artificial Bee Colony* (ABC) algorithm,[366] which more fully models the foraging dynamics of a real honeybee than the other algorithms, and is, arguably, the most widely used form of "bee swarm" intelligence. The ABC algorithm is defined using three types of bees (employed, onlookers and scouts), with one "employed" bee per food source. When employed bees come back to the hive from a source, they perform a dance. If the source runs dry, the bee turns into a scout and attempts to find another food source. Onlooker bees use the dances of employed bees to adjudicate selection of potential new food sources. The position of a food source denotes a possible solution to the optimization problem the ABC algorithm is being used to "solve," and the amount of nectar corresponds to the quality (i.e., fitness) of the associated solution.

As powerful as all SI-based methods are, and despite having been successfully applied to a wide variety of problems, it is important to note that no one technique represents a panacea solution to all types of problems. One finds that, in practice, certain problems are more amenable to this kind of solution scheme than others, and that it is not always a priori clear (from the nature of the "problem") why that is so. To emphasize: this is not merely a reasoned observation, but rather derives from the

[363] D. Karaboga and B. Akay, "A survey: algorithms simulating bee swarm intelligence," *Artificial Intelligence Review* 31, 2009.

[364] D. Pham, et al., "The bees algortihm: a novel tool for complex optimization problems," in *Proceedings of the 2nd Virtual International Conference on Intelligent Production Machines and Systems*, Cardiff, U.K., Elsevier, 2006.

[365] H. Wedde, et al., "BeeAdHoc," in Proceedings of the 2005 conference on Genetic and evolutionary computation: GECCO '05, 2005.

[366] D. Karaboga, "An Idea Based On Honey Bee Swarm for Numerical Optimization," Technical Report-TR06, *Erciyes University*, Computer Engineering Department 2005.

celebrated *No Free Lunch Theorem* (NFLT).[367] The theorem asserts that the performance of all search algorithms, when averaged over all possible cost functions (i.e., averaged over all possible problems), is *exactly the same*. In other words, no search algorithm is better or worse on average than blind guessing. To the extent that autonomy implies adaptation (to changing conditions, environments, unanticipated behaviors, etc.), and adaptation is, at root, an optimization process, the NFLT implies that there cannot exist a general strategy defined in some absolute way that will apply to every situation.[368] We will come back to this issue during a discussion of the risks and vulnerabilities associated with the design and deployment of autonomous systems, in general, and relying on bio-inspired computation in particular.

To the extent that an ability to engage in goal-directed behavior in unpredictable dynamic environments is a key attribute of autonomous systems, SI represents a powerful suite of tools for design, development, and analysis.

Big data

An everyday example of the power of "swarm intelligence" is the commercial exploitation of *Big Data*. "Big data" colloquially refers to the increasingly massive data storage capabilities, the proliferation of mobile networks, digital sensor networks, cloud computing, and cluster computer systems, which collectively are generating incomprehensibly massive worlds of information.[369] A recent report has found that, as of 2011, the total amount of digital content on the Internet is about 1.8 *zettabytes* (=1.8 $\times 10^{12}$ gigabytes), distributed among ~10^{17} separate files; and that figure is more than doubling every two years.[370] The opportunity to gain actionable insights from this data has already been recognized and increasingly relied on by businesses (mainly to tap into, and draw inferences from, user preferences, associations, and opinions about their products and services), and public and government agencies:

[367] D. Wolpert and W. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, 1996.

[368] R. Duro, J. Santos, and M Grana, *Biologically Inspired Robot Behavior Engineering*, Springer-Verlag, 2003.

[369] V. M.-Schonberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013.

[370] J. Gantz and D. Reinsel, *The 2011 Digital Universe Study: Extracting Value from Chaos*, IDC, June 2011: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

- Search engine companies such as *Google*, *Yahoo*, *Facebook*, and *Microsoft* thrive on the capture and proprietary analysis of the vast amount of freely available data on the World Wide Web (WWW). In 2012, *Google* released its *Knowledge Graph*,[371] a semantic search knowledge base; and, 2013, *Facebook* released its own semantic search engine that contains intimate knowledge about (its over one billion) users.[372]

- *Google* leverages its search data to estimate current flu activity (and predict outbreaks) around the world in near real-time.[373]

- *Netflix* (which has 27 million subscribers in the U.S. and 33 million worldwide) uses data-mined statistics and patterns of its viewers to both optimize its recommendations to customers of what they "would like" to watch,[374] and even help *define* original content to offer them.[375]

- *Walmart* and *Amazon*, apply machine learning algorithms to their warehouse of user-generated data to better manage their inventory and supply chains (e.g., *Walmart* archives 4 petabytes (4000 trillion bytes) worth of data that consists of every single purchase recorded by their point-of-sale terminals—around 267 million transactions per day—at their 6000 stores worldwide[376]).

- Credit card companies routinely scour over their reams of personal and financial information to identify patterns in consumer purchasing trends and detect fraud.[377]

[371] https://developers.google.com/knowledge-graph/.

[372] R. Golijan, "Facebook Graph Search," *NBC News*, July 9, 2013.

[373] A. Dugas, et.al., "Influenza Forecasting with Google Flu Trends," *PLOS One* 8, no. 2, 14 Feb 2013: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3572967/; *Flu Trends* homepage: http://www.google.org/flutrends/.

[374] D. Harris, "Netflix analyzes a lot of data about your viewing habits," Gigaom, 14 June 2012: http://gigaom.com/ 2012/06/14/netflix-analyzes-a-lot-of-data-about-your-viewing-habits/.

[375] D. Carr, "Giving Viewers What They Want," *New York Times*, 24 Feb 2013.

[376] R. Bryant, R. Katz and E. Lazowska, "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society," December 2008: http://www.cra.org/ccc/docs/init/Big_Data.pdf.

[377] K. Chaudhary, J. Yadav, and B. Mallick, "A review of Fraud Detection Techniques: Credit Card,"*International Journal of Computer Applications* 45, no. 1, May 2012.

- The advent of widespread use of social media has helped spawn an entirely new breed of commercial "big-data vendors," whose sole purpose is to collect and aggregate publicly available data (e.g., from forums, newsgroups, blogs, social networks, etc.) and to sell that data and/or their accompanying analysis, to help their clients (e.g., banks, car companies, retailers, etc.) exploit patterns in user preferences and behaviors, and craft better marketing strategies; e.g., the top three big-data vendors alone—*Splunk*, *Opera Solutions*, and *Mu Sigma*—generated $448 million in revenue in 2012, out of a total big-data market of $11.4 billion.[378]

- Public health officials now routinely parse *Twitter* feeds and search-engine query statistics to help identify the possible emergence of pandemics;[379] and psychologists are beginning to exploit SM to profile suspects in school shootings.[380]

Law-enforcement agencies, in particular—whose INTEL needs ostensibly differ from those of the military, but share certain key characteristics; e.g., answers to basic questions such as *Where?*, *When?*, *What?*, and *Who?*—have seizes a potential treasure trove of social-media-derived data in tracking criminals. For example, the New York Police Department stood up a unit in 2011 to track people who discuss their crimes on *Facebook*, *Twitter* and *MySpace*;[381] the Boston Police Department has used *Twitter* to monitor criminal activity in the city since 2009[382] (an effort that paid unexpected dividends during the recent bombing at the Boston Marathon[383]); and the FBI has used SM to investigate securities fraud and insider trading in the $2 trillion hedge fund industry.[384]

[378] K. Kelly et al., "Big Data Vendor Revenue and Market Forecast 2012-2017," *Wikibon*, 17 April 2013.

[379] A. Signorini, A.M. Segre, and P.M. Polgreen, 'The Use of Twitter to Track Levels of Disease Activity and Public Concern in the US During the Influenza A H1N1 Pandemic," *PLoS ONE* 6, no. 5, 2011.

[380] J. Hoffman, "Trying to Find a Cry of Desperation Amid the Facebook Drama," *New York Times*, 23 Feb 2012.

[381] R. Parascandola, "NYPD forms new social media unit to mine Facebook and Twitter for mayhem," *Daily News*, 10 August 2011.

[382] (1) http://www.cityofboston.gov/police/about/initiatives.asp; (2) http://www.twitter.com/boston_police.

[383] K. Bindley, "Boston Police Twitter: How Cop Team Tweets Led City From Terror To Joy," *Huffington Post*, 26 April, 2013.

[384] M. Goldstein and J. Ablan, "FBI uses Twitter, social media to look for securities fraud," *Reuters*, 26 Nov 2012.

Figure 21. Natural flocking of birds

upload.wikimedia.org/wikipedia/commons/d/d6/Fugle%2C_%C3%B8rns%C3%B8_073.jpg.

Rule-based flocking

One of the most breathtakingly beautiful displays of nature—and a prototype of self-organized emergence—is the synchronous, fluid like flocking of birds (see figure 21). Large or small, the magic of flocks is the impression they convey of some intentional centralized control directing the overall traffic. Though ornithologists still do not have a complete explanation for this phenomenon,[385] evidence strongly suggests that flocking is a decentralized activity, where each bird acts according to its local perceptions of what nearby birds are doing. Flocking is therefore a group behavior that emerges from collective action. (A popular form of computational "swarm intelligence," called *particle swarm optimization*,[386] is explicitly based on flocking.)

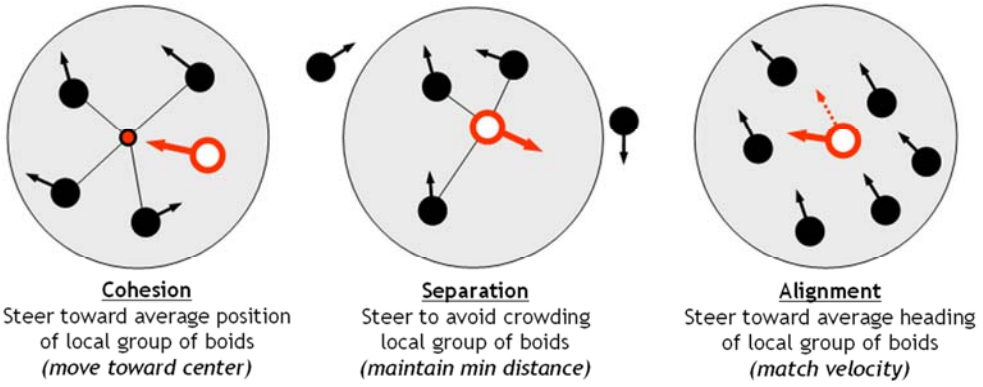
[385] P. Friederici, "Explaining Bird Flocks," *Audubon*, March-April 2009: http://www.audubon.org/magazine/march-april-2009/explaining-bird-flocks.

[386] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, Piscataway, N.J., IEEE, 1995.

In the 1980s, Craig Reynolds introduced a deceptively simple set of local behavioral "rules" which he programmed into a flock of artificial birds he called *Boids* (see figure 22):

1. Move toward the perceived center of nearby *Boids*
2. Maintain a minimum distance from other objects (including other *Boids*)
3. Match the velocity of nearby *Boids*

Figure 22. Three basic rules for "flocking"



Ref: C. Reynolds, "Flocks, herds, and schools: A distributed behavioral model," *Computer Graphics*, Vol. 21, No. 4, 1987.

Each *Boid* "sees" only what its neighbors are doing and acts accordingly. The collective motion of all the *Boids* was remarkably close to real flocking, despite the fact that there is nothing explicitly describing the flock as a whole. The *Boids* initially move rapidly together to form a flock. The *Boids* at the edges either slow down or speed up to maintain the flock's integrity. If the path bends or zigzags in any way, the *Boids* all make whatever minute adjustments need to be made to maintain the group structure. If the path is strewn with obstacles, the *Boids* flock around whatever is in their way naturally, sometimes temporarily splitting up to pass an obstacle before reassembling beyond it. There is no central command that dictates this action. The collective behavior is entirely unanticipated, and cannot be easily derived from the rules defining what each individual *Boid* does.

Reynolds' *Boids* rules (albeit along with a host of refinements) were used as the core set of local behaviors to drive the software-based "combat agents" in the ISAAC and EINSTein agent-based models of combat (developed by CNA in the late 1990s).[387]

[387] In addition to cohesion, separation, and alignment, ISAAC and EINSTein added a wide set of other primitive local behaviors (e.g., *find, fight, flee, follow path, pursue, evade, avoid,* and

These two models are among the earliest widely available simulation packages to apply "simple" flocking rules to military operations research analysis of combat dynamics. An AI-enhanced set of *Boids*-like rules was also used to build the Multiple Agent Simulation System in Virtual Environment (MASSIVE),[388] a crowd-simulation software package that was first used to generate battle scenes in the movie trilogy *The Lord of the Rings*.[389]

More recent examples that bridge the gap between software and hardware instantiation of rule-based flocking are: (1) autonomous drone flocks developed in 2011 at the Swiss Federal Institute of Technology in Lausanne,[390] and (2) a decentralized multi-copter flock developed in 2014 by a research team at the Eötvös University in Hungary.[391] While the first example is closer to being a prototype demonstration of "drone flocking" than a mature technology (it consists of fixed-wing fliers with no inter-drone interactions, and the drones can move only at constant speeds, and must fly at different heights to avoid collisions), the second example represents the first time that physical drones have actually flown as a resilient, coordinated dynamic flock.[392] Similar attempts had been made before, but always involved some constraints, such as restricting flights to controlled indoor environments or having the flock controlled by a central computer. In contrast, the drones developed by the Eötvös researchers involve no caveats: they can coordinate their movements to form rotating rings or straight lines, and if the drone-swarm is faced with, say, an obstacle such as a wall with a gap in it, it queues up to squeeze its way through just as Reynolds' software-based *Boids* did 30 years ago. The Eötvös-swarm also underscores the fact that instantiating in hardware otherwise "simple" flocking rules entails solving a number of engineering challenges; e.g., accounting for drift and noise in GPS signals (that the drones relied on for positioning), speeding up reaction times (so that drones do not approach so close that are miss or overshoot their marks), and problems associated with relying on radio for communication.

wander). Ref: A. Ilachinski, *Artificial War: Multiagent-Based Simulation of Combat*, World Scientific, 2004.

[388] http://www.massivesoftware.com/.

[389] D. Keoppel, "Massive Attack," *Popular Science Magazine* 261, no. 6, Dec 2002.

[390] S. Hauert, et al., "Reynolds flocking in reality with fixed-wing robots: communication range vs. maximum turning rate," *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 25-30, 2011.

[391] G. Vásárhelyi, "Outdoor flocking and formation flight with autonomous aerial robots," presented at the *IEEE IROS Conference*, 2014: https://arxiv.org/ftp/arxiv/papers/1402/1402.3588.pdf.

[392] E. Yong, "Autonomous drones flock like birds," *Nature*, 26 Feb 2014.

Cooperative tasking

An early example of an emergent "group mind" behavior that spontaneously appears without being centrally orchestrated—and is among the earliest examples cooperative tasking by robot teams—is a decentralized sorting algorithm introduced in the 1990s.[393] Inspired by the self-organized manner in which real ant colonies sort their brood, the algorithm has simple robots move about a fenced-in environment that is randomly littered with objects that can be scooped up. These robots: (1) *move randomly*, (2) *do not communicate with each other*, (3) *can perceive only those objects directly in front of them* (but can distinguish between two or more types of objects with some degree of error), and (4) *do not obey any centralized control*. The probability that a robot picks up or puts down an object is a function of the number of the same objects that it has encountered in the past.

Coordinated by the positive feedback these simple rules induce between robots and their environment, the result, over time, is (a seemingly intelligent) coordinated sorting activity. Clusters of randomly distributed objects spontaneously and quite naturally emerge out of a simple set of autonomous local actions having nothing at all to do with clustering per se. The system's simplicity, flexibility, error tolerance, and reliability compensates for their lower efficiency. While one can argue that this collective sorting algorithm is much less efficient than a hierarchical one, the cost of having a hierarchy is that the sorting would no longer be ant-like but would require a god-like oracle analyzing how many objects of what type are where, deciding how best to communicate strategy to the ants. Furthermore, the ants would require some sort of internal map, a rudimentary intelligence to deal with fluctuations and surprises in the environment (what if an object was not where the oracle said it would be?), and so on. In short, a hierarchy, while potentially more efficient, would of necessity have to be considerably more complex as well. A much simpler collective decentralized system can lead to seemingly intelligent behavior while being more flexible, more tolerant of errors, and more reliable than a hierarchical system.

A more recent example of robotic-teams cooperatively building structures without a centralized controller is the TERMES system, developed at Harvard in 2014.[394]

[393] R. Beckers, E. Holland, and J. Deneubourg, "From local actions to global tasks: stigmergy and collective robotics," pages 181-189 in *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, edited by R. Brooks and P. Maes, MIT Press, 1994.

[394] J. Werfel, K. Petersen, and R. Nagpal, "Designing Collective Behavior in a Termite-Inspired Robot Construction Team," *Science* 343, no. 6172, 14 Feb 2014. A short video of the construction process is also available: E. Gibney, "Termite-inspired robots build castles,"

Inspired by the natural "engineering" abilities of termites, TERMES robots use foam bricks to build towers, castles, and pyramids, autonomously, adding bricks wherever they are needed, and even creating staircases in order to reach higher levels of their buildings. As with most hardware instantiations of software-driven *Boids*-like rules, the local behaviors are both few in number and simple; e.g., (1) move forward, backward, and turn in place; (2) climb up or down a step the height of one brick; (3) pick up a brick, carry it, and deposit it directly in front of itself; (4) Detect other bricks and robots in immediate vicinity; and (5) keep track of its own location with respect to a "seed" brick. The robots themselves are also simple, possessing four basic types of sensors and three actuators. Structures arise—or, more precisely, *emerge*—via stigmergy; i.e., the aforementioned form of implicit communication within a swarm in which individual robots observe each other's changes to the environment and act accordingly.

The TERMES system is a proof-of-concept for robust scalable distributed swarm intelligence. Each robot executes its actions in parallel with the entire swarm, and without knowing what other robots are working on at the same time. Swarms are robust, in that the overall construction process continues unabated in the event that one or a few robots either malfunctions or, for whatever reason, fails to accomplish its local goal. And TERMES is scalable, since exactly the same rule set can be executed by 5 robots, 50, 500, or more. (A related effort by the same research team behind TERMES has also demonstrated self-assembly of a *thousand-robot* swarm, in which individual micro-bots are able to assemble into complex preprogrammed shapes). [395]

There are two "takeaways" from these examples: (1) "swarm intelligence" represents a powerful general method by which self-organized emergence (of complex adaptive systems) and distributed cooperative problem solving (through self-organization) can be used to design autonomous robotic swarms, and (2) despite the simplicity and age of the basic rule sets on which they based (e.g., the *Boids* rules were introduced in 1987, and "decentralized sorting" dates back to 1994), essentially the same mechanisms can, and are, being used to "program" many of today's state-of-the-art robotic swarms. This entails both *benefits* (e.g., the research community can rely on several decades' worth of experience in designing and exploring the behaviors of "software swarms") and *challenges* (e.g., essentially the same fundamental problems and questions that were asked when the first "software swarm" prototypes were being designed apply equally to their hardware-instantiated brethren; and not all such problems have been completely "solved").

Nature News, 3 Feb 2014: http://www.nature.com/news/termite-inspired-robots-build-castles-1.14713.

[395] M. Rubenstein, A. Cornejo, R. Nagpal, "Programmable self-assembly in a thousand-robot swarm," *Science* 354, no. 6198, 15 Aug 2014.

Engineering robotic swarms

Among the most pressing questions for current robotic swarm technology is, *"Can swarms be designed?"*; i.e., can a specific set of desired emergent behaviors be "programmed" by specifying a set of local behavioral rules? Before answering this question, we first draw a suggestive analogy between the recent emergence of robotic swarms with that of neural networks (NNs) over the last 20 to 30 years.

In the case of NNs (as per our earlier discussion), recent successes such as *AlphaGo*'s prowess in Go[396] derive not from some radically new approach—the basic principles behind NNs (if not deep learning techniques, since even this most recent incarnation is, loosely speaking, an extension of the core NN model)—but from the fact that computers have finally become fast enough and able to quickly access enough memory to instantiate what was heretofore a "proof of concept"-level-technology. The most recent development of Deep learning NN systems has also been greatly accelerated by using massively parallel Graphics Processing Units (GPUs) for training.[397] And there has been an increasing drive toward a direct hardware implementation of artificial neural networks patterned after the human brain (via so-called "neuromorphic circuits").[398] The takeaway is that all of these advances are, at their root, latter-day realizations of general techniques introduced decades earlier.

The recent growth and development of robotic swarms has followed an analogous evolutionary path; i.e., from early conceptualization to modern implementation. It is also one that—arguably, even more so than for NNs—is ripe with directly applicable "lessons learned" from several decades' worth of multidisciplinary research in various interrelated fields such as *complex adaptive systems* (CASs),[399] *artificial life*,[400] and *multi-agent-based modeling*.[401]

[396] S. Byford, "Google's AlphaGo AI beats Lee Se-Dol again to win Go series 4-1," *The Verge*, March 15, 2016.

[397] S. Jones, "NVIDIA GPUs Power Deep-Learning Winners in World Cup of Image Recognition," *NVidia*, 7 Sep 2014: https://blogs.nvidia.com/blog/2014/09/07/imagenet/.

[398] K. Ramanaiah and S. Sridhar, "Hardware Implementation of Artificial Neural Networks," *i-manager's Journal on Embedded Systems* 3, no. 4, Nov 2014 - Jan 2015.

[399] J. Miller and S. Page, *Complex Adaptive Systems*, Princeton University Press, 2007.

[400] C. Langton, *Artificial Life: An Overview*, MIT Press, 1997.

[401] S. Railsback and V. Grimm, *Agent-Based and Individual-Based Modeling: A Practical Introduction*, Princeton University Press, 2011.

Many of the rules and behaviors that are now being instantiated in individual robots and robotic swarms were first developed to support software-based studies of complex systems. Because the focus of early research (in the 1980s and 1990s) was mainly to achieve a basic understanding of the general behavior of complex systems—which, by their nature, are not easily amenable to traditionally reductionist forms of mathematical analyses—simulations were the primary tool for most studies. Out of the artificial life community (spawned in concert with the founding of the *Santa Fe Institute* (SFI) in 1984,[402] which remains one of the world's premier centers for CAS research), emerged a powerful new set of simulation methods—called *multi-agent based models* (MBMs)—designed specifically for studying the dynamics of complex systems, in general, and swarms, in particular.

Early MBMs were developed not to design behaviors, but to *understand them.*[403] They were developed to help answer basic question such as "How, and under what conditions, do specific behaviors arise?", "Which behaviors are unique to a given system, and which are generalizable?", and "Are there universal sets of behaviors?" While MBMs continue to serve as key "go to" exploratory probes of basic CAS behaviors,[404] the robotic swarm industry has moved away from "MBMs = simulations as distillations" (to gain insight) to "MBMs = simulation-based rules and algorithms as *descriptions*" of real (i.e., engineered) robots and behaviors. It is here, at the cusp between exploring behaviors and prescribing rules that generate them, that MBMs—and modeling and simulation, in general—can help mitigate some of the challenges and uncertainties of developing autonomous systems and robotic swarms.

Turning our attention back to the question posed earlier, *"Can swarms be designed?"*… Since swarms represent a particular class of self-organized emergent behaviors that arise in complex adaptive systems, they are not generally amenable to conventional "design" processes.[405] Indeed, as of this writing (Nov 2016), *no general method exists that maps individual rules to (desired) group behavior.*[406] Nonetheless, "swarm engineering" methods exist to facilitate the unique design requirements of robotic swarms; though each has its own concomitant benefits, limitations, and

[402] http://www.santafe.edu/about/the-history/founders/.

[403] The first widely available general purpose MBM modeling environment was developed at SFI, and was called SWARM http://www.swarm.org/wiki/Main_Page.

[404] *Journal of Artificial Societies and Social Simulation*: http://jasss.soc.surrey.ac.uk/JASSS.html.

[405] M. Prokopenko, "Design versus self-organization," Chapter 1 in *Advances in Applied Self-Organizing Systems*, Second Edition, edited by M. Prokopenko, Springer-Verlag, 2013.

[406] I. Navarro and F. Matia, "An Introduction to Swarm Robotics," *International Scholarly Research Notes* 2013, 2013: https://www.hindawi.com/ journals/isrn/2013/608164/.

challenges;[407] one common challenge is Verification and Validation (V&V), and will be discussed later. All design methods fall into one of two general classes:[408] (1) *behavior-based design*, and (2) *automatic design*.

The simplest, and most common, behavior-based design approach to programming desired behavior is bottom-up, trial-and-error. That is, iteratively adjusting and tuning individual rules until the resulting collective behavior is achieved.[409] The local rules themselves can assume a variety of interrelated forms; e.g., *finite-state machines* (in which specific local states are mapped to local actions via, say, a probabilistic response threshold function),[410] *virtual physics* (in which individual robots behave, locally, as virtual particles that exert attractive and repulsive forces on other robots),[411] and *property-driven design* (in which a set of desired collective behaviors are first described as logical propositions, and the local rules consistent with that top-level description are iteratively defined).[412]

Automatic design methods—i.e., those that do not explicitly require the direct intervention of the developer—include:[413] *reinforcement learning* (RL), in which local rules are "taught," or self-learned, via trial-and-error interactions with the environment that provide positive and negative feedback,[414] and *evolutionary*

[407] "Swarm engineering," as a robotic design methodology, was introduced, as a concept by Kazadi in 2000 (S. Kazadi, *Swarm Engineering*, PhD thesis, California Institute of Technology, Pasadeba, CA), and more formally in a seminal paper by Winfield, et al. in 2004 (A. Winfield,

C. Harper, and J. Nembrini, "Towards dependable swarms and a new discipline of swarm engineering," in *Proceedings of the International Workshop on Simulation of Adaptive Behavior* 3342 of Lecture Notes in Computer Science, Springer-Verlag).

[408] V. Gazi, B. Fidan, L. Marques, and R. Ordonez, "Robot Swarms: Dynamics and Control," in *Mobile Robots for Dynamic Environments*, edited by E. Kececi and M. Ceccarelli, Momentum Press, 2015.

[409] V. Crespi, A. Galstyan, and K. Lerman, "Top-down vs bottom-up methodologies in multi-agent system design," *Autonomous Robots* 24, no. 3, 2008.

[410] E. Bonabeau, et al., "Adaptive task allocation inspired by a model of division of labor in social insects, in *Biocomputing and Emergent Computation: Proceedings of BCEC97*, World Scientific Press, 1997.

[411] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *The International Journal of Robotics Research* 5, no. 1, 1986.

[412] M. Brambilla, et al., "Property-driven design for swarm robotics," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, edited by L. Conitzer, et al., International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[413] E. Kececi and M. Ceccarelli, editors, *Mobile Robots for Dynamic Environments*, Momentum Press, 2015.

[414] M. Kochenderfer, et al., *Decision Making Under Uncertainty*, MIT Press, 2015.

programming (EP), in which the dynamics of natural Darwinian selection and evolution are applied directly to *breeding* desired local and collective behaviors.[415] The main drawback to both methods is the computational cost incurred by searching through a vast space of possible behaviors and the complexity of robot-robot interactions. RL also suffers from the so-called "credit assignment" problem, which refers to need to identify and distribute the overall reward for "group behavior" among individual robots (thereby defining their local rules). Applications of EP to robotic design are also plagued by the fact that convergence to solutions (i.e., a set of local rules that gives rise to a desired set of global behaviors) is not guaranteed.[416] Additional problems with both automatic design methods include (1) the a priori difficulty of accounting for all possible environmental states and perceptions of states, which—during the design phase, will necessarily be incomplete—or even knowing when an "accounting" is sufficiently complete),[417] and (2) dealing with dynamic environments (i.e., accounting for actions of other robots responding to changes the collective makes to the environment itself). Both issues remain "open" research areas in the design of robotic swarms.

Controlling robotic swarms

A second pressing question for robotic swarm technology is, *"How can swarms be controlled (or supervised)?"* Whatever are the means by which a swarm is engineered (as discussed in the previous section), there is the further issue of ensuring that a deployed swarm successfully accomplishes whatever set of tasks it is assigned. While the supervision of multiple vehicles builds on how they are individually programmed, actual control has not yet achieved a "plug and play" simplicity. For one thing, it requires the operator(s) to have an intimate familiarity with and an understanding of robot behavior (as semi-autonomous entities). For another, the operator(s) must be facile with the communication protocols and general interface between whatever primitives are under her own control and whatever consequent actions are induced on the part of individual robots. Since the latter, in turn, depends on the degree to which an individual member of a swarm is autonomous[418]—which is generally a function not just of an agent's innate properties, but also reflects an

[415] S. Nolfi and D. Floreano, *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*, MIT Press, 2000.

[416] M. Brambilla, et al., *Swarm robotics*, IRIDIA Technical Report Series, R/IRIDIA/2012-014.

[417] L. Kaelbling, M. Littman, and A. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence* 101, 1998.

[418] Autonomy is formally discussed in the next section of this report.

agent's dynamic context—there is an additional challenge of controlling or supervising a swarm with less than complete information (about robots' local environments).

Although there is extensive research on how to *design* swarms (e.g., via multiagent-based modeling techniques; see pages []-[]), and control *individual* robots, the literature on human supervisory control of multiple semi-autonomous robots is more sparse; much of it remains cutting edge. Just as there is currently no general method that maps rules that describe context-dependent actions of individual robots to desired group behaviors (see previous section), there are (as of this writing) *no validated schemes for scalable, flexible, and adaptive human control of robot teams.*[419]

The earliest research into a single operator controlling multiple robots appeared during the middle to late 1990s in Ph.D. dissertations.[420] Since then, the majority of Human-Robot Interaction (HRI) studies takes place under the auspices of traditional human factors domains, including psychology, industrial engineering, and aeronautical engineering. Presentations on single-human control of multiple robots—including unmanned ground vehicles (UGVs), unmanned aerial vehicles (UAVs), and unmanned underwater vehicles (UUVs)—have appeared regularly in the past half-decade at major robotics conferences (e.g., IEEE Conference on Robots and Automation,[421] IEEE/RSJ Conference on Intelligent Robots and Systems,[422] and IEEE Symposium on Robots and Human Interactive Communication[423]).

Figure 23 shows, schematically, the key components of a human-swarm system: the *human operator* (for whom a set of "cognitive complexity" functions is displayed on the left; and the meaning of which is discussed below), the *swarm* (highlighted in blue on the right, and consisting of multiple robots, $\{A_1, \ldots, A_m\}$, some or all of whom may be linked by radio and/or stigmergic interactions via environment), and the *human-swarm interface*, shown in the center, through which the human supervises and/or controls the swarm, and the swarm informs the operator of its current state and behavior.

[419] K. Sycara, *Robust Human Interaction with Robotic Swarms*, presentation at ICAART 2016: https://vimeo.com/161163755.

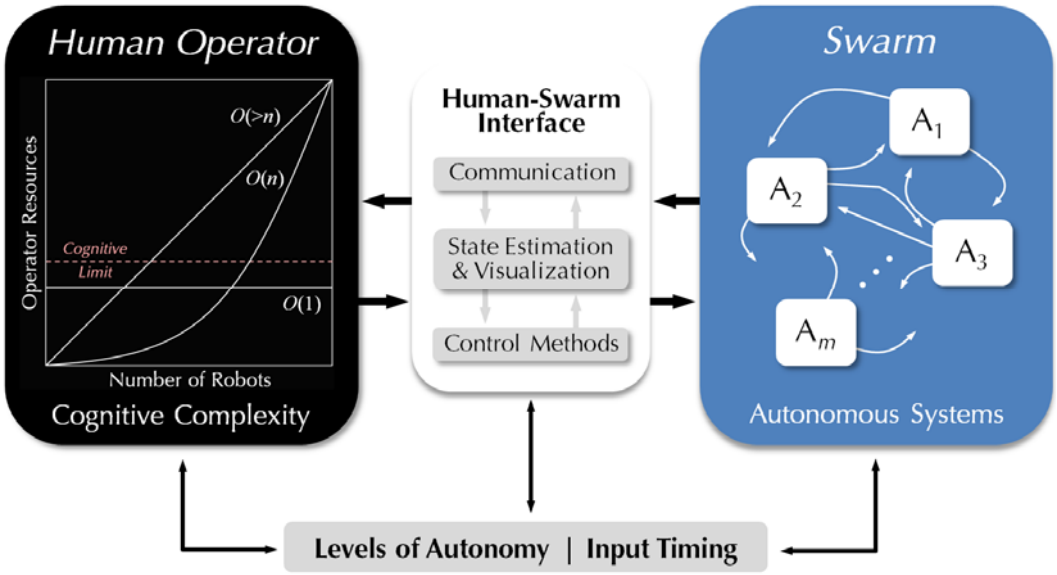
[420] J. Adams, *Human management of a hierarchical system for the control of multiple mobile robots*, Ph.D. Dissertation, University of Pennsylvania, Philadelphia, 1995; K. Ali, Multiagent telerobotics: Matching systems to Tasks, Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, 1999.

[421] https://www.icra2016.org.

[422] http://www.iros2016.org.

[423] http://ro-man2016.org.

Figure 23. Key components of human-swarm behavior and control



After A. Kolling, et al., "Human Interaction with Robot Swarms," *IEEE Transactions on Human-Machine Systems*, Vol. 46, Feb. 2016.

Taxonomies of multi-robot systems focus mainly on physical characteristics, tasking, and methods; while taxonomies of human-robot interaction (HRI) include roles and structure. For example, one early taxonomy that has been refined through the years uses seven factors to describe robot teams based on size, three communication-related dimensions, reconfigurability, processing ability, and heterogeneity.[424] Others have used essentially the same classification, but have emphasized the ability of a robot (within a swarm) to be aware of, and coordinate with, other robots.[425] HRI taxonomies include those by Scholtz, et al.,[426] which classifies the human role according to type of control exerted over the swarm (e.g., supervisor, operator,

[424] G. Dudek, M. Jenkin, E. Milios, and D. Wilkes, "A taxonomy for swarm robots," in *Proceedings of the 1993 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 1993; G. Dudek, M. Jenkin, and E.Milios, "A taxonomy of multirobot systems," in *Robot Teams: From Diversity to Polymorphism*, edited by T. Balch and L. Parker A. K. Peters, 2002.

[425] A. Farinelli, L. Iocchi, and D. Nardi, "Multirobot systems: A classification focused on coordination," *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 34, 2004.

[426] J. Scholtz, M. Theofanos, and B. Antonishek, "Theory and evaluation of human robot interactions," in *36th Hawaii International Conference on Systems Sciences*, IEEE, 2002.

mechanic, peer, or bystander); Yanco and Drury,[427] which adds the possibility of interactions with swarms (factors include the numbers of humans and robots, and their possible links for communication and/or coordination); and Goodrich and Schultz,[428] which distinguishes between remote and proximate interactions.

Cognitive complexity

A more recent HRI-based taxonomy of multi-robot systems, introduced by Lewis in 2013,[429] is based on ranking the difficulty of the operator's *tasks*. Lewis introduces the concept of "cognitive complexity" (an offshoot of "computational complexity," which is a measure used in computer science to describe the time it takes to solve a problem as a function of the size of its input)[430] to describe the relationship between task types and command complexity. The operator's mental workload (defined in a classic text on human factors engineering as "a measurable quantity of the information processing demands placed on an individual by a task")[431] has been shown to be an important factor in determining the number of robots an operator can control.[432]

The control of multiple robots is analogous to the execution of a computer algorithm in that both are defined by patterns of sequences of state-dependent decisions and actions. For example, if an operator is tasked with controlling or supervising n robots, each of whom executes its actions independently of other robots, the operator can devote an equal amount of attention to each robot, in turn; therefore his cognitive complexity is "of order n," written $O(n)$. This indicates that the required effort on the part of the operator devoted to controlling the swarm scales linearly with the number of the robots (e.g., search and rescue scenarios). Moreover, since operator actions do not interfere with one another for linearly scaled systems, the number of required controllers also scales linearly with the size of the swarm.

[427] H. Yanco and J. Drury, "Classifying human–robot interaction: An updated taxonomy," *Proceedings of the IEEE Conference on Systems, Man and Cybernetics* 3, 2002.

[428] M. Goodrich and A. Schultz, "Human–robot interaction: A survey," *Foundations and Trends in Human–Computer Interaction*, Vol. 1, 2007.

[429] M. Lewis, "Human interaction with multiple remote robots," *Reviews of Human Factors and Ergonomics* 9, no. 1, 2013.

[430] C. Moore and S. Mertens, *The Nature of Computation*, Oxford University Press, 2011.

[431] M. Sanders and E. McCormick, *Human Factors in Engineering and Design*, Mcgraw-Hill, 1993.

[432] D. Olsen and S. Wood, "Fan-out: Measuring human control of multiple robots," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2004.

If a swarm can be tasked with a single command (e.g., by designating an area to be reconnoitered, or—more generally—where the members of the swarm are capable of coordinating autonomously, such as by flocking or rendezvous), the operator's cognitive complexity will be independent of the number of robots; i.e., the cognitive complexity is fixed; or is $O(1)$. $O(1)$ tasks implicitly assume that individual robots are fully autonomous, a constant demand being placed only on the human operator. $O(1)$ control thus describes situations in which there are a large number of robots that can be coordinated using relatively simple goals (e.g., following a predefined plan of action). Bio-inspired swarms fall into this class.

On the other hand, if the actions of one robot are dependent on the states and/or actions of other robots in the swarm, the operator's cognitive complexity is nonlinear; i.e., it is of order $O(>n)$. Such tasks cannot be specified simply and, depending on the details of robot-robot interactions, may require arbitrarily complex controls on the part of the operator. For example, a study of human control and coordination of box-pushing robots (via assignment of waypoints and monitoring formation following) found that the operators become rapidly saturated, with little time remaining for assigning additional tasks.[433]

Of course, in practice, it is unlikely that swarms will be controlled at any fixed level of complexity. Rather, control over individual robots is likely to be distributed over different people (during various parts of a scenario and/or simultaneously) at different levels of complexity.[434] "Cognitive complexity," as a metric, is not intended as a panacea description of human-robot control; rather, its purpose is to merely focus attention on the level of effort required of human controllers to interact with multi-robotic systems, including swarms. Indeed, a complementary set of metrics describing "levels of autonomy" are discussed in the next section (and shown to be equally wanting in terms of completeness and utility), where it will be argued that a major current gap in the development of autonomous systems is a lack of a comprehensive conceptual framework in which to organize the multiple simultaneous dimensions of unmanned systems (including, but not limited to, the human-system control interface).

An additional challenge to coordinating the actions of swarms (is getting the timing right (see bottom of figure 23). That is, making sure that the timing of—and between— commands issued to individual robots of a swarm are commensurate with both their internal clocks (as defined by the "flocking algorithm") and environment

[433] G. Kaminka, and Y. Elmaliach, "Single operator, multiple robots: Call-request handling in tight coordination tasks," *Distributed Autonomous Robotic Systems* 7, 2006.

[434] A. Kolling, et al., "Human-swarm interaction: an experimental study of two types of intercation with foraging swarms," *Journal of Human-Robot Interaction* 2, no. 2, 2013.

(i.e. as dictated by the temporal rhythms of physical changes). The same command(s) issued to the same robot(s) at different times and/or in different dynamic contexts may lead to unpredictable and/or undesirable behavior. For example, a basic problem that arises in all flocking algorithms is to understand the conditions under which a swarm may fragment,[435] and which commands are appropriate to issue at what time to help the swarm regain its cohesion.

The three basic components of the human-swarm interface are (see central panel in figure 23):[436] (1) *communication*, (2) *state estimation and visualization*, and (3) *control methods*. Optimizing communication between the human operator (who typically resides at a remote terminal, relative to the position and area of operations of the swarm) and individual robots is a key challenge for robotic swarm technology in general. Specifically, the problem is to convolve (the nature and behavior of) fundamentally decentralized distributed systems with some form of centralized control. Part of the challenge is of a conventional nature, in that it involves solving the same basic latency, bandwidth, and asynchrony issues that typically arise in traditional "non swarming" networked systems, though there are few studies dedicated to understanding how these issues impact swarm dynamics.[437] Other challenges are unique to robotic swarms, and stem from myriad uncertainties associated with "less than perfect" physical realizations of mathematical idealizations. By design (since swarms are typically "engineered" to display desired behaviors using mathematical distillations; e.g., agent-based models), autonomous robots are assumed to perform perfectly without human intervention; and their behavioral profiles are predicated on this assumption holding true in real-world environments. Of course, real robots are guaranteed to behave imperfectly. Communications will not always be reliable (or possible at all, such as for underwater systems), aspects of a robot's environment (which may be critical for human

[435] For example, S. Nagavalli, L. Luo, and K. Sycara ("Neglect benevolence in human control of robotic swarms," in *International Conference on Robotics and Automation (ICRA)*, IEEE, 2014) show that improper timing of control input could lead to swarm fragmentation.

[436] A. Kolling et al., "Human Interaction with Robot Swarms: A Survey," *IEEE Transactions on Human-Machine Systems* 46, Feb. 2016.

[437] Generally speaking, bandwidth is more limited and latency and asynchrony are both higher in swarms than in other types of systems. One study found that an increase in latency led to deteriorating performance in a foraging swarm—though if the operator had the ability to predict behavior, the negative effects of latency could be ameliorated (P. Walker et al., "Neglect benevolence in human control of swarms in the presence of latency," in *IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, 2012). There is an irony in that many of the most powerful methods used to "solve" these otherwise conventional technical challenges are themselves derived from swarm-based optimization methods; e.g., M. Saleem, G. Di Caro, and M. Farooq, "Swarm intelligence based routing protocol for wireless sensor networks: Survey and future directions," *Information Sciences* 181, 2011.

intervention or control) will not always be known (a robot may veer off course due to wind or other unforeseen random element), and, depending on the form of data that a robot is designed to communicate back to the operator, there may not even be enough time for the human to assimilate the requisite information to enact a desired change in a swarm's behavior (i.e., the operator's cognitive complexity may exceed her limit).

Hayes and Adams[438] provide numerous other examples of *physica*l-state (e.g., position, speed, direction, group membership and local clustering), and *virtual*-state (e.g., the ID of which robot is the "leader" of the swarm, which may also change as a scenario unfolds) uncertainties, drawing parallels between known impacts on the behavior of biological swarms and the potential implications such uncertainties may have for robotic swarms.

State estimation and visualization (middle of central panel in figure 23) refers to the need for an operator to be able to observe the state and evolution of a swarm, as well as predict its likely future states. More precisely, the operator must understand the possible dynamic impact(s) that specific "controls" may have on the swarm's behavior. Of course, difficulties immediately arise when the information available to the operator is less than complete and/or is provided in a less-than-timely fashion. Kolling at al.,[439] provide a survey of recent studies that explore the impact of constraints of bandwidth, latency, and display design on the operator's ability to visualize a swarm. Not surprisingly, many of the same methods used to *design* swarms may also be used to *predict* and *control* their behavior.

Methods of control

There are four general (and partly overlapping) approaches to controlling multiple robots, all predicated on the supposition that the operator's cognitive complexity will scale as $O(1)$; i.e., that the swarm can be viewed as a single entity:[440]

1. *Behavior–based*, in which the human controller uses a palette of primitive behaviors that each "agent" of the swarm is endowed with, and that it can perform autonomously.

[438] S. Hayes and J. Adams, "Human-swarm interaction: sources of uncertainty," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, ACM, 2014.

[439] A. Kolling, et al., "Human Interaction with Robot Swarms: A Survey," *IEEE Transactions on Human-Machine Systems* 46, Feb. 2016.

[440] G. Coppin and F. Legras, "Autonomy Spectrum and Performance Perception Issues in Swarm Supervisory Control," *Proceedings of the IEEE* 100, no. 3, March 2012.

2. *Policy-based*, in which the human operator issues group goals and/or constraints on (expected) behavior, leaving the individual members of the swarm to choose their own course of action; rather than direct each robot.

3. *Playbook*-based, in which the human operator issues a specific plan of action from a master "playbook" (analogous to book of plays used by players on a football team), that can be defined on different levels of granularity, depending on tactical context.

4. *Proxy-agent-based*, in which a software interface is used as an intermediary between the human operator and individual agents of the swarm; one that communicates (and negotiates) on behalf of the swarm.

Behavior–based approaches are arguably the most straightforward to apply, in the sense that the main element involved, namely the set of rules that the individual agents are programmed to follow—e.g., loiter, move forward, and rally at a given position— are all fixed beforehand and are known to the human operator.

The basic method is to augment an agent's innate primitive "flocking rules" (see figure 22) with rules that infuse an operator's influence. Typical "control" rules include:[441] *leading* (by which the human operator needs to manage only one swarm member, with other agents "following" by virtual attraction), acting as *predator* (similar to leading but in which swarm members are repelled by the leader rather than attracted; this seemingly counterintuitive method of control can more readily split swarms into groups), and control via *stakeholders* (or "special" agents, whereby certain privileged members of a swarm are controlled by the human operator but are not otherwise recognized as "different" by other members of the swarm; such methods have been found to useful in selectively guiding the collective behavior of the swarm[442]). Control can also be enacted at various levels:[443] (1) switching between specific algorithms that implement desired behaviors, (2) changing the value of selected parameters of a given algorithm, (3) indirectly altering a swarm's behavior by selectively changing features of the environment.

In principle, operators can have at their disposal as large a library of algorithms implementing specific swarm behaviors as deemed necessary to execute a given mission. And behavior-based control generally works best when the robots have a

[441] B. Pendleton and M. Goodrich, "Scalable Human Interaction with Robotic Swarms," *AIAA Infotech@Aerospace (I@A) Conference*, Boston, MA, 2013

[442] S. Kerman, D. Brown, and M. Goodrich, "Supporting human interaction with robust robot swarms," *5th IEEE International Symposium on Resilient Control Systems*, IEEE, 2012.

[443] A. Kolling, et al., "Human Interaction with Robot Swarms: A Survey," *IEEE Transactions on Human-Machine Systems* 46, Feb. 2016.

sufficiently high degree of autonomy (some metrics for which are introduced in the next section); i.e., they are able to perform their tasks with minimal error and minimal human oversight (between "controls").

In all of this, the operator's challenge is to understand (and be able to anticipate the consequences of) those rules well enough to be confident that the agents will perform desired actions, and that the swarm will perform as desired as a whole. Of course, as outlined in the previous section, neither of these outcomes is a given; nor is it a given that the operator will be able to monitor the progress of a swarm in uninterrupted fashion. Whether a swarm is controlled by switching between different algorithms or making selective changes to parameter values (regulating an otherwise fixed algorithm), there is an irreducible uncertainty in the effect any change—however small—will have on the swarm's overall behavior; effects of changes emerge from the interactions within individual robots making up the swarm, and the environment.

Behavior–based approaches are most useful in cases for which a small number of rules are sufficient to cover basic operational needs (for simple scenarios). They become increasingly difficult to apply as the complexity of a mission scales up, and/or the number of required robots increases (for which the ability to predict and "micro-manage" group behavior rapidly becomes infeasible).[444]

Policy-based approaches have the virtue that they spare the human operators the "complication" of issuing orders to individual agents of a swarm. Of course, the presumption is that the policies have been engineered well enough beforehand to ensure that the swarm, as a whole, behaves as desired. The policies themselves may assume a variety of forms; e.g., notification, delegation, supervision, or constraints. The "devil is in the details" resides in designing an appropriate "language" that describes policy-orders in a way that is precise (read: mathematical) enough to yield unambiguous swarm behavior, yet is "simple" enough to understood by a human operator who may not be programming savvy. Specific technologies supporting such an interface are usually text or speech based (the latter of which also involves aspects of natural language processing), and require a detailed understanding of autonomy. A major challenge is to make sure that the interface between human-operator-issued policies and swarm behaviors is able to continuously and robustly

[444] M. Wilson and M. Neal, "Diminishing returns of engineering effort in telerobotic systems," *IEEE Trans. Syst. Man Cybern. A*, Syst. Humans 31, *Special Issue on Socially Intelligent Agents: The Human in the Loop*, no. 5, Sep. 2001.

adjust the degree of autonomy that is appropriate for whatever dynamic context the swarm happens to be in at a given time.[445]

Playbook-based approaches[446] rely on tailoring actions (selected from a "master set" of pre-defined plans-of-action and/or tactics to use) to dynamic contexts, giving the operator a range of autonomous behaviors to use in various situations. The idea is for operators to call out "plays" that then trigger desired patterns of behavior. Since all members of the swarm "know" the same playbook and "understand" what their (and other agents') roles are for all actions, regulating their behavior ought—in principle—to entail less of a communications burden than some other approaches to controlling the swarm.

[445] J. Bradshaw et al., "Dimensions of Adjustable Autonomy and Mixed-Initiative Interaction," in *Agents and Computational Autonomy* 2969 of the series *Lecture Notes in Computer Science*, Springer-Verlag, 2004.

[446] C. Miller and R. Parasuraman, "Designing for flexible interaction between humans and automation: delegation interfaces for supervisory control," *Human Factors* 49, 2007.

Autonomy

Although "autonomy" is nowadays a seemingly ubiquitous concept, lying at the heart of most discussions (including this report) that pertain to AI and robotics, in general, and unmanned systems, in particular, it is notoriously difficult to define (examples of this "difficulty" appear throughout this section). And yet, given its growing importance to—indeed, given its key role in shaping DoD's future acquisition of—modern weapon systems, it is a concept that needs to be understood *precisely*. Yet, as mentioned earlier, even the otherwise laudably comprehensive Defense Science Board's most recent report on autonomy provides but a cursory definition.[447]

Etymologically, the word comes from Greek, and is a fusion of *autos* (meaning "self") and *nomos* (meaning "law"); meaning, "self-governing." But the word's meaning has, over the centuries, been muddled by its diffused appropriation by multiple disciplines. For example, in the 18th century, Kant argued that autonomy is to understood as a moral action consistent with one's free will;[448] and Piaget later suggested that autonomy—i.e., the ability to self-govern—is a critical component of child development.[449]

More recently, autonomy has entered the lexicon of robotics, where it—as a concept—straddles an ambiguous middle-ground between sets of properties that define a robot's "human like" qualities (including analogues of a "moral code" that may underlie decisions) and a robots engineering-level characteristics and capabilities. That the meaning of the word remains muddled owes itself mostly to the fact that the use of the term reflects the multi-disciplinary nature of the field.[450] Since robotics itself is a complex amalgam of engineering, computer science, cognitive science, and artificial intelligence, there is (as yet) no definitive universally-

[447] *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Sec. of Def. for Acquisition, Technology and Logistics, June 2016.

[448] Manuel Kant, *Correspondence*, Cambridge University Press, 2007.

[449] Jean Piaget, *The Moral Judgment of the Child*, Free Press, 1997.

[450] J. Beer, A. Fisk, and W. Rogers, "Toward a Psychological Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction* 3, 2014.

agreed on definition of autonomy. Indeed, there are references to all of the following:[451]

> Adaptive autonomy, adjustable autonomy, agent autonomy, basic autonomy, behavioral autonomy, belief autonomy, biological autonomy, causal autonomy, constitutive autonomy, energy autonomy, mental autonomy, motivational autonomy, norm autonomy, robotic autonomy, shared autonomy, sliding autonomy, social autonomy, subservient autonomy, user autonomy, among many others.

"Autonomy" is also often confused with "automatic." Automatic systems refer to simple systems that functions with no (or limited) human operator involvement, typically in structured and unchanging environments, and whose performance is limited to the specific set of actions (usually, well-defined tasks that have predetermined "scripted" responses).

More complex automatic systems are sometimes referred to as "automated," and are usually defined as rule-based dynamical systems (e.g., self-driving cars, and many military weapon systems fall into this category). Neither "automatic" nor "automated" systems require external control or guidance. And, even as "autonomy," as a term, is (as yet) ill-defined as an operational concept, it is also often confused with general "intelligence"; i.e., with systems that are capable of human-level cognition and understanding. And, even here, there is added confusion between "intelligent" systems that do well on specific problems (referred to before as "narrow AI"; e.g., IBM's computer chess AI *Deep Blue*) and those that mimic human ability across multiple problem domains ("general AI"). Unfortunately, there are no objective boundaries between any of these categories.

Further complicating the issue is that a system's innate complexity (however it is defined!) is independent of: (1) the type/degree of human control to which it is tied, (2) the tasks and missions it performs, and (3) the complexity of the environment in which it must execute its mission.

All of this can (and does) lead to serious semantic confusion, as a system may be "autonomous" in the sense that it operates entirely without human intervention, but otherwise functions relatively simply, leading some to describe it as "automatic" or "automated."[452] An example of a framework that attempts to account for this multidimensional aspect of autonomy is discussed later in this section.

---

[451] D. Vernon, *Artificial Cognitive Systems: A Primer*, MIT Press, 2014; L.G. Shattuck, "Transitioning to Autonomy: A human systems integration perspective," Presentation at *Transitioning to Autonomy: Changes in the role of humans in air transportation*, March 11, 2015: https://humanfactors.arc.nasa.gov/workshop/autonomy/download/presentations Shaddock%20.pdf.

[452] P. Scharre, *Autonomous Weapons and Operational Risk,* Center for a New American Security, Feb 2016.

Figure 24. A sampling of definitions of "autonomy" and "autonomous systems"

Definitions of "Autonomy" and "autonomous systems" (or autonomous agents/robots)	
"The state of existing or **acting separately** from others." [1]	"…the condition or quality of being **self-governing**." [8]
Autonomous "…agents operate **without the direct intervention** of humans or others, and have some kind of control over their actions and internal states." [2]	"A capability (or a set of capabilities) that enables a particular action of a system to be automatic or, within programmed boundaries, '**self-governing**.'" [9]
"An autonomous agent is a system situated within and a part of an environment that sense that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future;" "Exercises **control over its own actions**." [3]	Autonomous systems "…have a set of intelligence-based capabilities that allow it to respond to situations that were not pre-programmed or anticipated in the design. Autonomous systems have a degree of **self-government** and **self-directed behavior**" [10]
"The robot should be able to carry out its actions and to refine or **modify the task and its own behavior** according to the current goal and execution context of its task." [4]	"An autonomous system is **self-directed** by choosing the behavior it follows to reach a human-directed goal.." [11]
An autonomous robot …can operate, self-contained, under all reasonable conditions without requiring recourse to a human operator. Autonomy means that a robot can **adapt to change in its environment** … or itself … and continue to reach a goal." [5]	A weapon system that, once activated, **can select and engage targets without further intervention by a human operator**…includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation … but can select and engage targets without further human input after activation." [12]
"An unmanned system's own ability of sensing, perceiving, analyzing, communicating, planning, decision-making, and acting, to achieve goals as assigned by its human operator(s) through designed HRI … The condition or quality of being **self-governing**." [6]	Able to "… **independently compose and select among different courses of action** to accomplish goals based on its knowledge and understanding of the world, itself, and the situation." [13]
"… ability to **accommodate variations in its environment**. Different robots exhibit different degrees of autonomy; the degree of autonomy is often measured by relating the degree at which the environment can be varied to the mean time between failures, and other factors indicative of robot performance." [7]	"Autonomy is … the degree to which the system has the capability to **achieve mission goals independently**, performing well under significant uncertainties, for extended periods of time, with limited or non-existent communication, and with the ability to compensate for system failures, all without external intervention." [14]

[1] *Merriam-Webster on-line dictionary*: http://www.merriam-webster.com/.
[2] M. Wooldridge and N. Jennings, "Intelligent agents: theory and practice," *The Knowledge Engineering Review*, Vol. 10, 1995.
[3] S. Franklin and A. Graesser, "A Taxonomy for Autonomous Agents," in *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, Springer-Verlag, 1996.
[4] R. Alami, et al., "An Architecture for Autonomy," *Inter. Jour. of Robotics Res.*17, 1998.
[5] R. Murphy, *An Introduction to AI Robotics*, MIT Press, 2000.
[6] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson , 2009.
[7] S. Thrun, "Toward a framework for human-robot interaction," *Human-Computer Interaction*, 2004.
[8] *Autonomy Levels for Unmanned Systems (ALFUS) Framework*, *Volume I : Terminology*, Version 2.0, National Institute of Standards and Technology, Special Publication 1011-I-2.0, October 2008.
[9] *The Role of Autonomy in DoD Systems*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.
[10] *Technology Investment Strategy: 2015-2018*, Autonomy Community of Interest (COI), TEVV Working Group, ASD(R&E) , May 2015.
[11] *Unmanned Systems Integrated Roadmap: FY2013-2038*, U.S. Department of Defense.
[12] DoD Directive 3000.09, *Autonomy in Weapon Systems*, Nov 2012.
[13] *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016.
[14] *Autonomous Horizons: System Autonomy in the Air Force, A Path to the Future*, *Volume I: Human-Autonomy Teaming*, U.S. Air Force, Office of the Chief Scientist, June 2015.

Figure 24 shows a small sampling of definitions of "autonomy" and "autonomous systems," culled from both the academic research community and recent DoD memos and reports (references are provided at the bottom of the figure). Highlighted

in red are the most commonly appearing fragments and phrases: *act separately… self-governing… without direct intervention… control over own actions… self-directed behavior… modify own behavior… adapt to changes in environment… independently compose and select actions… accommodate variations in environment… select / engage targets without intervention by human operator… achieve mission goals independently*. (We note in passing that, as of this writing, Nov 2016, the most current on-line version of DoD's dictionary of military terms does not include any entries on either "autonomy" or "autonomous systems."[453])

The Defense Science Board's recent study on autonomy[454] defines an autonomous system as one that is able to independently compose and adjudicate among a set of possible actions to accomplish goals based on its knowledge and understanding of the world and itself, and able to adapt to dynamic contexts in its environment.

Before amplifying on this definition, and discussing ways of moving beyond definitions to articulating differences among *levels of autonomy*, and toward developing a conceptual framework that embraces the multidimensional aspects of autonomy, we pause briefly to summarize the potential operational benefits of autonomy (however it is defined).

Operational benefits of autonomy

There are a number of ways in which (varying degrees of) autonomous capabilities may potentially benefit military operations. Some extend and complement human performance, some provide direct replacements of humans (in parts of the loop), and still others may portend entirely new operational capabilities:[455]

[453] Though it does provide definitions for an "unmanned system" (="an aircraft that does not carry a human operator and is capable of flight with or without human remote control") and "unmanned aircraft system" (="[a] system whose components include the necessary equipment, network, and personnel to control an unmanned aircraft"), along with an entry for (human based) "autonomous operations." Ref: JP 1-02, *Department of Defense Dictionary of Military and Associated Terms*, 8 November 2010 (as amended through 15 Feb 2016): http://www.dtic.mil/doctrine/new_pubs/jp1_02.pdf.

[454] *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016; http://www.acq.osd.mil/dsb/reports/DSBSS15.pdf.

[455] G. Zacharias, *Autonomous Horizons: System Autonomy in the Air Force*, Presentation at CogSIMA 2016, San Diego, CA, 24 March 2016; *Technical Assessment: Autonomy*, Department of Defense, Office of Technical Intelligence, Office of the Assistant Secretary of Defense for Research and Engineering, Feb 2015; *Summer Study on Autonomy*, Department of Defense,

- *Reduced costs:* depending on mission domain, autonomy may potentially reduce system cost (e.g., long duration and continuous-operation data-acquisition-intensive missions, which traditionally require great manpower, self-diagnosing autonomous systems would reduce maintenance costs); the caveat is that since DoD has only recently started to seriously consider autonomy-related issues, details of the true life-cycle costs involved are unclear. Since autonomy will involve a considerable technological investment, may require greater overall manning, and will need networks with significantly higher capacity than existing systems, life-cycle costs may just as likely increase as they are to decrease. For example, *Predator*, *Global Hawk*, and many other operational UAVs may all require a minimal set of human operators to *fly*, but need a significantly larger support staff for planning, maintenance, analysis, etc.[456]
- *Reduced risk (of human injury/death):* autonomous systems reduce the number of humans required to operate in dangerous areas, thereby fundamentally reducing risk (e.g., contested operations, route clearance and mine sweeping, chemical, biological, and/or radiological environments).
- *Freedom from certain human limitations:* workload, fatigue, stress, emotions (anger, fear, etc.); longer flight times for unmanned aerial vehicles and the ability to loiter in larger geographic areas.
- *Reduced risk (of cyber-attack):* since remotely piloted unmanned systems typically rely on a satellite tether to a human pilot and are unable to complete their mission if the communication link is severed, the mission of an autonomous system can potentially continue unimpeded even in heavily contested A2/AD environments. Enable operations with denied or degraded communication links.
- *Increased persistence and endurance:* autonomy increases both mission duration (e.g., enabling unmanned vehicles) and persistent surveillance.
- *Mission expansion:* autonomy introduces the potential to tap into heretofore unexplored new types of operations and CONOPS (e.g., heterogeneous autonomous swarms).
- *Enhanced mission performance in time:* autonomy effectively both extends and shortens timescales, encroaching on, and far transcending, extremes of human ability (e.g., missile defense, long duration ISR acquisition and

Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016.

[456] P. J. Springer, *Military Robots and Drones*, ABS-CLIO, 2013.

analysis, cyber operations); it is likely that only autonomous systems will be able to keep pace with the increasingly face tempo of warfare.[457]

- *Enhanced assimilation/understanding of data:* greater autonomy affords increased ability to cope with high volume data and diversity of data types (e.g., imagery, intelligence data analysis, ISR data fusion). Synchronization of activities of platforms, software, and operators over wider scopes and ranges (e.g., manned↔unmanned-system teaming).

- *Reduced "mission space" complexity:* autonomy militates against the increasing—and otherwise potentially paralyzing—complexity in multimodal decision making situations (e.g., a Combined Air Operations Center (CAOC), multi-mission operations).

- *Mitigation of data loss:* autonomy militates against intermittent and or denied communications in contested environments and/or undersea operations.

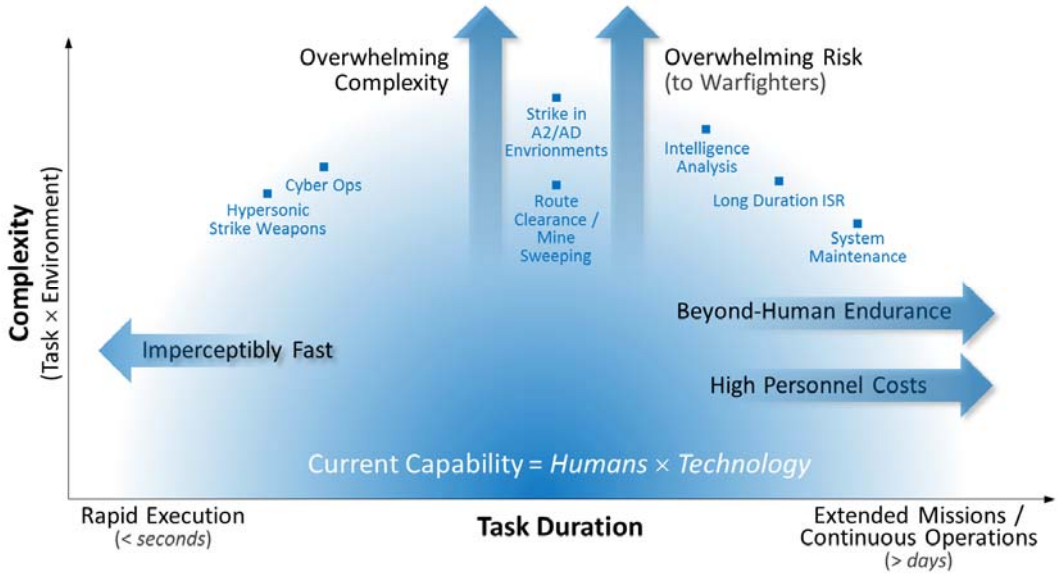
Figure 25 illustrates the challenges to existing human-machine systems and potential benefits and opportunities for autonomy. The inverted "U" (highlighted in blue) denotes, notionally, the current limit of human-machine performance as viewed in the context of *task duration* (*x*-axis) and *complexity* (*y*-axis = complexity of task × environment). Current limits are imposed by physical bounds on human performance, technology, research and development budgets, and ethics. The arrows depict some of the key challenges to existing capabilities; e.g., imperceptibly fast execution of tasks, extended mission timelines that strain human endurance, and increasing complexity of mission environments.

The elements highlighted in blue in figure 25 depict mission areas in which autonomy can potentially enhance performance, reduce risk, and/or reduce costs over what is currently possible, For example, to the extent that cyber operations and hypersonic strike weapon warfare already represent mission spaces that proceed on time-scales far too short to allow for meaningful human decision-making to intervene, autonomy is indispensable. At the other extreme of the "task duration" spectrum, autonomy would allow otherwise interminably long missions (for humans)—such as Intelligence, Surveillance, and Reconnaissance (ISR) operations—to

[457] P. W. Singer, Director of the 21st Century Defense Initiative at the Brookings Institute, suggests in his book, *Wired for War* (Penguin, 2009) that the human location "in the loop" is already becoming that of a "supervisor who serves in a fail-safe capacity in the event of a system malfunction," and that the speed, confusion, and information overload of modern combat will soon push the process outside of "human space" altogether. Future weapons, "will be too fast, too small, too numerous, and will create an environment too complex for humans to direct."

proceed virtually continuously. Unburdened by the nominal 12-hour limit of a human in the cockpit, autonomy would allow sensors and precision weapons to be placed in areas of interest at greater distances for longer periods of time than now possible, thereby enhancing situational awareness to all levels of command.

Figure 25. Challenges to existing human-machine systems and opportunities for autonomous capabilities



Combines figures 1 – 4 in: *Technical Assessment: Autonomy*, DoD, Office of Technical Intelligence, Office of the Assistant Secretary of Defense for Research and Engineering, Feb 2015

In addition to enhanced performance at the extremes of task durations, autonomy can also help militate against the increasing complexity of the mission space itself, such as in A2/AD environments (albeit, only if the autonomous system(s) are equipped with the most sophisticated AI, since—to be of any real help to the warfighter/mission-commander—such systems must be able to process, and derive inferences from, a litany of raw data and changing conditions: from integrated air-defense systems, to jamming, to mobile intelligent targets). At the very least, even in less-than-extreme complex environments, autonomy can help mitigate risk to manned systems by eliminating the need for human presence. This is already done, to a limited extent, by use of unmanned ground vehicles to reconnoiter potential

IEDs, although—due to current limits on image-recognition and cognitive abilities—the actual disposal still requires a manned presence).[458]

Domain-specific capabilities

The Defense Science Board's 2012 Task Force report on autonomy[459] and RAND's recent study on designing unmanned systems with greater autonomy[460] summarize the current capabilities and potential benefits of autonomy in four operational domains:

- *Unmanned Aerial Vehicles* (UAVs): UAVs are already capable of performing a variety of missions to support both the military and intelligence communities, from ISR (which has also been integrated with strike on the same unmanned platform), to over the horizon targeting, anti-ship missile defense, ship classification, electronic warfare and signals intelligence, to deception operations, to direct connectivity of UAV operators to ground forces.[461] Existing technology permits autonomous landing capability using on-board sensors (though some information regarding the landing site must still be preloaded);[462] which, when refined, can help reduce the number of human operators needed to operate a fleet of UAVs. Notably, due to developments in sense-and-avoid technologies, the accident rate for most unmanned systems is now essentially that of manned aircraft. The concept of Remote-Split Operations (RSO), in which the human control of UAVs in multiple locations can be switched between controlling aircraft in different theaters as mission and weather requirements dictate and conduct shift changes in mid-flight, was introduced by the Air Force in 2003.[463]

[458] A. Amanatiadis, et al., "The AVERT project: Autonomous Vehicle Emergency Recovery Tool," in *Robotics and Automation (ICRA) IEEE International Conference*, 2015.

[459] Section 2 in *The Role of Autonomy in DoD Systems*, DoD Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

[460] D. Gonzales and S. Harting, *Designing Unmanned Systems with Greater Autonomy*, Rand Corporation, 2014.

[461] S. Gupta, M. Ghonge, and P. Jawandhiya, "Review of Unmanned Aircraft System (UAS), *Inter. Journal of Advanced Research in Comp. Eng. and Tech* 2, no. 4, April 2013.

[462] Paul Williams and Michael Crump, "Intelligent Landing System for Landing UAVs at Unsurveyed Airfields," *Proceedings of the 28th Inter. Congress of the Aeronautical Sciences, 2012:* http://icas.org/ICAS_ARCHIVE/ICAS2012/PAPERS/131.PDF.

[463] Megan Orton, "General Underscores Commitment to Fielding Unmanned Aerial Systems," *American Forces Press Service*, 14 January 2009.

Current gaps include: (1) no high-fidelity training environments for UAV pilots; (2) no computer-based training system for *Predator* crews to operate in conjunction with real-world weapons tactics training; and (3) no full simulation training system exist to ensure that the level of proficiency of aerial unmanned crews is maintained. Moreover, the different military services take vastly different approaches to training (e.g., the Air Force requires ten months to fully train a Predator crew member, but the Army requires only three months).

- *Unmanned Ground Vehicles* (UGVs): the key benefit of UGVs is similar to that of the UAV; namely, they provide a persistent standoff capability. They are currently deployed mostly in support of counter-IED and route clearance operations, using robotic arms attached to, and operated by, modified Mine Resistant Ambush Protected (MRAP) vehicles and remotely controlled robotic systems; to a lesser extent, UGVs are used for reconnaissance in dismounted and tactical operations.[464]

  The key challenges for incorporating autonomy in UGVs include: (1) dynamic terrain negotiation and obstacle avoidance, and (2) performing kinetic operations within the Rules of Engagement (ROE). The first challenge requires basic image and pattern recognition skills; while the second requires a more sophisticated capacity to reason (e.g., the system must be able to make quick "on the fly" decisions that are both consistent with the ROE and reflect the changing conditions of the mission space). Current state-of-the-art UGVs are designed to operate only within a well-defined environment. As long as the conditions are consistent with that environment, the UGVs perform as expected. However, when conditions are "sufficiently different" (the drivers for which are impossible to exhaustively *pre*-test for), the UGV's behavior will no longer be predictable. The UGV must be "smart enough" to perceive, understand, and adapt to the changing dynamic contexts of its environment.[465]

  DARPA was a pioneer in the development of autonomous ground vehicles, holding the first of its "grand challenges" in 2004 (with others following in later years; the most recent one was held in 2012 DARPA).[466] The challenges

[464] "Mine Resistant Ambush Protected (MRAP) Armored Vehicles," *Defense Update*: http://defense-update.com/products/m/mrap.htm.

[465] *Unmanned Systems Integrated Roadmap: FY2013-2038*, Under Secretary of Defense, Acquisition, Technology and Logistics, Reference Number 14-S-0553, Washington, DC: Department of Defense, 2013.

[466] Special Issue on the DARPA Grand Challenge, Part 1, *Journal of Field Robotics* 23, issue 8, August 2006.

are essentially races in simple urban and rural environments, in which competing teams must field systems that can navigate a racecourse filled with various obstacles, and be first to reach the finish line. In recent years, of course, every major automaker is developing (and/or deploying) autonomous vehicle technologies.[467] Notably, the "amount of research and development funds going into civilian autonomous vehicle development will likely greatly exceed that available for UGV R&D in the DoD budget over the next decade."[468]

- *Unmanned Maritime Vehicles* (UMVs): which is a category that encompasses both surface (USVs) and underwater (UUVs) vehicles. UMVs may be used for ensuring security within harbors, scanning for problems on a ship hull, sweep for mines, secure critical waterways, and provide ocean tracking. Unlike UAVs (which cannot operate in bad weather or low visibility), UMVs can operate in poor weather conditions. However, although persistence is a key capability, since water attenuates radio waves and other wireless signals, maintaining communication with UUVs (and even with surface vehicles) is technically challenging, particularly at longer ranges. Although there are technologies available to militate against this fundamental limitation (e.g., the use of laser communication systems), they are generally expensive and require massive amounts of power.[469] Thus, there is an essential need for UUVs to have an autonomous ability to plan and execute their own paths, and to avoid obstacles and other unanticipated underwater terrain elements.

  UUVs already have varying levels of autonomous capabilities. Examples include:[470] GPS/Doppler-aided navigation; autonomous path planning and execution based on onboard world map; terrain-following, and keep-out zone avoidance; autonomous decision making and cue generation for noncombat missions; dynamic replanning based on sensor input (acoustic, radio frequency (RF), chemical, etc.), vehicle health, and mission objectives and priorities; cross-deck advanced autonomy on multiple classes of vehicles (interface to various vehicle controllers and payload controllers).

[467] J. Anderson et al., *Autonomous Vehicle Technology: A Guide for Policymakers*, RAND Corporation, RR-443-1-RC, 2014.

[468] Page 30 in D. Gonzales and S. Harting, *Designing Unmanned Systems with Greater Autonomy*, Rand Corporation, 2014.

[469] M. Scholz, "Using Laser Communication Above Water and Underwater," *Sea Technology Magazine,* 2011: https://www.sea-technology.com/features/2011/0511/laser_communication.php.

[470] D. Ashton (Capt), "Unmanned Maritime Systems Autonomy," presentation at *10th International MIW Technology Symposium*, 7-10 May 2012.

Examples of technologies permitting higher levels of autonomy that are currently being developed include:[471] long transit and autonomous planning and control to precise local insertion without GPS-aided navigation (i.e., bottom map-matching/feature-based navigation); adaptive area surveys with automated target detection, classification, and recognition; robust sense and avoidance of hard-to-image/classify obstacles (e.g., surface vessel detection and avoidance, threat avoidance, and RF spectrum threat counterdetection); autonomous sensor data fusion; collaborative behaviors; and fault detection and response.

Capabilities still out of reach for UUVs include many of the same limitations that currently apply to other unmanned systems, including: counterdetection awareness and response, real-time sensor processing, dynamic threat perception and adversary intent, autonomous decision making to support the use of weapons, and advanced collaborative behaviors.

USV missions include:[472] antisubmarine warfare (ASW), maritime security, surface warfare, special operations forces support, electronic warfare and maritime interdiction operations support.

USV challenges include: autonomous implementation of the international regulations for preventing collisions at sea,[473] mine detection and classification, surface threat detection and perception, tracking and targeting a moving surface target (e.g., high-speed swarm), detection and tracking of submerged targets, autonomous target lock for kinetic action, and decision making for targeting and weapon release.

While the challenges and opportunities outlined above are all well understood by the defense community, there is as yet no overarching universally agreed upon conceptual framework in which the myriad tradeoffs can be objectively assesse; the pathway toward which is discussed next (starting with DoD's basic definition of autonomy).

[471] Ibid.

[472] *The Navy Unmanned Surface Vehicle Master Plan*, July 2007: http://www.navy.mil/navydata/technology/usvmppr.pdf.

[473] *COLREGS: International Regulations for Preventing Collisions at Sea*, Articles of the Convention on the International Regulations for Preventing Collisions at Sea, 1972, International Maritime Organization, 2005.

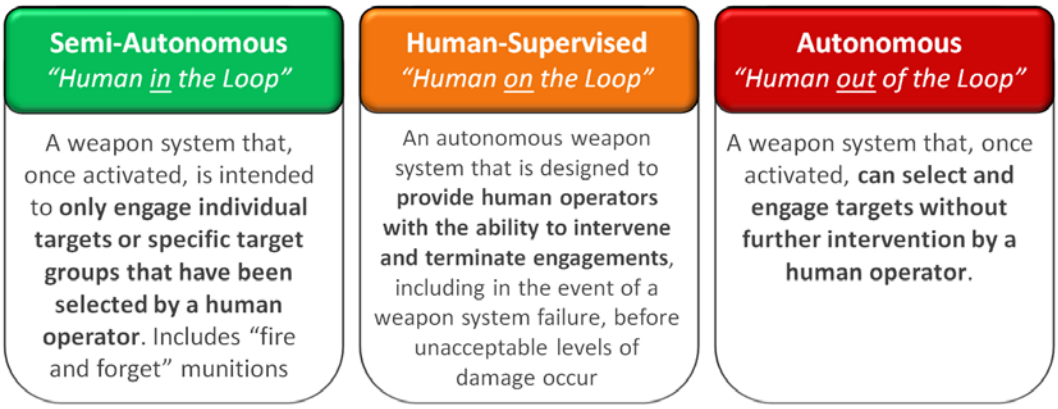
DoD's current definition of *autonomy*

DoD's Directive (DoDD) 3000.09 (*Autonomy in Weapon Systems*) establishes policy, organizational responsibilities, and "…guidelines designed to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements."[474]

The directive defines three (partly overlapping)[475] categories of autonomy in terms of the degree to which a human is involved in an autonomous system's performance (see figure 26):

- *Semi-autonomous*: human "in the loop"
- *Human-supervised:* human "on the loop"
- *Autonomous:* human "out of the loop"

Figure 26. Definitions of various levels of autonomy that appear in DoD 3000.09



[474] Under Secretary of Defense for Policy, DoD Directive 3000.09, *Autonomy in Weapon Systems*, Washington, DC, Nov 2012: http://www.dtic.mil/whs/directives/ corres/pdf/300009p.pdf

[475] The possibility for overlap is intentional, and allows for the fact that a given system may have subsystems that operate in different modes of autonomy during different phases of a particular mission. Also, DoD distinguishes between autonomy and remote control: "When the aircraft is under remote control, it is not autonomous. And when it is autonomous, it is not under remote control." (*Unmanned Systems Integrated Roadmap: FY2013-2038*, Under Secretary of Defense, Acquisition, Technology and Logistics, Reference Number 14-S-0553, Washington, DC: Department of Defense, 2013, p. 15.)

Human "in the loop"

In a *semi*-autonomous system, the machine stops and waits for human approval before continuing after each task is completed. The human operator is assumed able to monitor the environment and the machine's actions, and gives a "go ahead" to the machine once it has been confirmed that the machine's performance is adequate and consistent with operational mission requirements.

Human operators can play any of three essential roles (sometimes simultaneously) in terms of target selection and engagement:[476]

- As *essential operators:* in which the weapon system cannot effectively complete engagements without the human operator.
- As *moral agents*: in which the human operator makes value-based judgments regarding the use of force; e.g., weighing the probability of destroying a military target versus potentially incurring collateral damage.
- As *fail-safes*: in which the human operator has the ability to intervene and either change or halt the weapon system's operation in the event that the weapon malfunctions or conditions no longer warrant an engagement.

This hybrid "human-machine teaming" has recently been christened "Centaur Warfighting,"[477] after the half-human, half-horse creatures in Greek mythology.[478] An example of the power of this approach has recently been demonstrated in the world of chess, with Gary Kasparov's founding of the field of "advanced chess"[479]—essentially a form of Centaur chess—in which human chess-players and AI-chess playing systems cooperatively compete against opponents; humans still make the final moves, but they are based on a vastly richer base of information: "[the] idea was to create the highest level of chess ever played, a synthesis of the best of man and machine…[humans can] concentrate on strategic planning instead of spending so much time on calculations." Kasparov recounts how, in a chess tournament in 2005 that welcomed all players (amateurs and professionals) and allowed the use of

[476] P. Scharre, *Autonomous Weapons and Operational Risk,* Center for a New American Security, Feb 2016.

[477] S. Freedberg, Jr., "Centaur Army: Bob Work, Robotics, & The Third Offset Strategy," *Breaking Defense*, 9 Nov 2015.

[478] http://www.greekmythology.com/Myths/Creatures/Centaur/centaur.html.

[479] G. Kasparov, "The Chess Master and the Computer," *The New York Review of Books*, 11 Feb 2010: http://www.nybooks.com/articles/2010/02/11/the-chess-master-and-the-computer/.

computers, two noteworthy results stood out: (1) teams of human plus machine dominated even the strongest computers (a single *Deep Blue* caliber chess-AI player lost badly to a strong human player using a relatively weak laptop), and (2) the winner of the tournament was *not* a grandmaster using a state-of-the-art AI program, but rather two amateur chess players using three "ordinary" computers.

> "Their skill at manipulating and "coaching" their computers to look very deeply into positions effectively counteracted the superior chess understanding of their grandmaster opponents and the greater computational power of other participants. Weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process."[480]

Human "on the loop"

In a *supervised*-autonomous system, once activated, the machine performs a task under human supervision, and will continue performing the task until the human operator intervenes to halt its operation. However, in practice, there will always be a time delay between when a malfunction or failure occurs and when the operator exerts whatever "control" is necessary to adjust the machine's behavior. For example, it may take some time for the human operator to simply "recognize" (and/or understand the reasons for) a malfunction, and there may be some delay in deciding on what commands are appropriate to send to the machine to correct its behavior.

Examples of human-supervised autonomous weapon systems include:[481]

- **Drones**
  - *X-47B:* Northrop Grumman's fighter-size drone prototype; designed for autonomous launch and landing capability on aircraft carriers and able to navigate autonomously
  - *Taranis:* U.K.'s combat drone prototype; designed to autonomously search, identify and locate targets, but allowed to engage target only when authorized by mission command (also has autonomous self-defend capability)
  - *Harpy:*[482] Israel Defense Forces (IDF's) "fire-and-forget" AWS; designed to detect, attack and destroy radar emitters

[480] Ibid.

[481] Appendix B in *An Introduction to Autonomy in Weapons Systems*, P. Scharre and M. Horowitz, Center for a New American Security, Feb 2015.

- **Air/Missile Defense Systems**
  - *Aegis Combat System:*[483] Centralized, automated, command-and-control (C2) and weapons control system
  - *Goalkeeper:*[484] Dutch Close-In Weapon System (CIWS); deployed with a number of operators including the Royal Navy, Belgian Navy and South Korean Navy
  - *Iron Dome:*[485] IDF's air defense system, deployed since 2011
  - *Kashtan:*[486] Russian CIWS provides defense against anti-ship missiles, anti-radar missiles and guided bombs
  - *Mk-15 Phalanx CIWS:*[487] Fast-reaction, detect-through-engage, radar guided, 20-millimeter gun weapon system; C-RAM (Counter Rocket, Artillery and Mortar system) is essentially a land-based equivalent[488]
  - *Patriot:*[489] Long-range, all-altitude, all-weather air defense system to counter tactical ballistic missiles, cruise missiles and advanced aircraft
- **Ground Robot Active Protection Systems (APSs)**
  - AMAP-ADS:[490] German APS; also known as *AAC* in Sweden and as *Shark* in France; modular design can be adapted to a broad range of vehicles

[482] P. Spielman, "Israeli killer robots could be banned under UN proposal," *The Times of Israel*, 3 May 2013.

[483] *Aegis Weapon System*, United States Navy fact File: http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=200&ct=2.

[484] "Thales to upgrade Goalkeeper weapon system for Dutch Navy," Naval-technology.com, 3 Dec 2012: http://www.naval-technology.com/news/newsthales-upgrade-goalkeeper-weapon-system-dutch-navy.

[485] R. Wootliff, "Israel successfully tests shipborne Iron Dome missile interceptor," *The Times of Israel*, 18 May 2016.

[486]http://www.navyrecognition.com/index.php/east-european-navies-vessels-ships-equipment/russian-navy-vessels-ships-equipment/weapons-a-systems/123-kashtan-kashtan-m-kashtan-lr-cads-n-1-close-in-weapon-system-ciws-.html.

[487] *Mk-15 Phalanx CIWS*, United States Navy fact File: http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2.

[488] https://www.msl.army.mil/Pages/C-RAM/default.html.

[489] *Patriot Missile Long-Range Air-Defence System*, Army-technology.com: http://www.army-technology.com/projects/patriot/.

[490] B. Dodson, "Rheinmetall tests new Active Defense System under live fire," *New Atlas*, 1 Feb 2012: http://newatlas.com/rheinmetall-ads-live-fire-test/21278/.

- ARENA:[491] Russian APS designed to protect armored fighting vehicles from destruction by light anti-tank weapons, anti-tank guided missiles (ATGM), and missiles with top attack warheads
- DROZD (Thrush)/DROZD-2: Russian tank active protection system originally installed on T-55 and T-62 series main battle tanks (MBTs), and more recently (2005+) installed on T-62, T-72, T-80, T-90 types.
- *Iron Curtain:*[492] originated as a DARPA program (2005); uses high-speed sensing and parallel processing to intercept and destroy a multitude of threats inches from their targets
- *SGR-A1:*[493] Samsung's military robot sentry; deployed in DMZ between South/North Korea, autonomously detects targets, semi-autonomous engagements (but, reportedly, equipped with "full auto" mode)
- SWORDS:[494] the Special Weapons Observation Reconnaissance Detection System (SWORDS) robot that can carry lethal weaponry (M240 or M249 machine guns, or a .50 caliber rifle); three were used in in Iraq and Afghanistan. A new Modular Advanced Armed Robotic System (MAARS) version is in development.[495]
- *Trophy*:[496] IDF's APS intercepts and destroys incoming missiles and rockets with a shotgun-like blast; also known as ASPRO-A

As of this writing (Nov 2016), at least 30 nations use some form of supervised autonomous defensive systems in which humans are "on the loop" for selecting and engaging specific targets.[497] However, to date, these systems have been used only defensively, and to target objects (e.g., missiles, rockets or aircraft), not people.

[491] A. Geibel, "Learning from their mistakes: Russia's Arena Active Protection System," *ARMOR magazine*, 1 Sep 1996.

[492] http://artisllc.com/iron_curtain_active_protection_system/.

[493] http://www.globalsecurity.org/military/world/rok/sgr-a1.htm.

[494] "The Inside Story of the SWORDS Armed Robot "Pullout" in Iraq," *Popular Mechanics*, 30 Oct, 2009: http://www.popularmechanics.com/technology/gadgets/a2804/4258963/.

[495] https://www.qinetiq-na.com/products/unmanned-systems/maars/.

[496]Trophy Active Protection System, *Defense Update*: http://www.defense-update.com/products/t/trophy.htm.

[497] Australia, Bahrain, Belgium, Canada, Chile, China, Egypt, France, Germany, Greece, India, Israel, Japan, Kuwait, the Netherlands, New Zealand, Norway, Pakistan, Poland, Portugal, Qatar, Russia, Saudi Arabia, South Africa, South Korea, Spain, Taiwan, the United Arab Emirates, the United Kingdom, and the United States. Ref: P. Scharre and M. Horowitz, *An Introduction to Autonomy in Weapons Systems*, Center for a New American Security, Feb 2015.

Human "out of the loop"

In fully-autonomous systems, once activated, the machine performs its tasks without any assistance on the part of the human operator, who neither supervises the operation nor has an ability to intervene in the event of a system failure.

To date, there have been few human "out of the loop" autonomous weapon systems that select and engage its own targets. One example is the class of *loitering attack munitions* (LAMs).[498] LAMs are cruise missile-like devices that are launched into a general area and whose mission is to loiter, looking for targets according to pre-programmed targeting criteria (e.g., enemy radars, ships or tanks); once a target is detected, the LAM will fly into the target to destroy it. The only currently operational LAM is the Israel Defense Forces (IDF's) *Harpy*, a "fire-and-forget" anti-radar weapon that flies a general search pattern over a designated area to search for enemy radars, which, if one is found, then dive-bombs into it to destroy it. Examples of experimental LAMs that were not operationally deployed include the low-cost autonomous attack system (LOCAAS),[499] designed to target tanks, and *Tacit Rainbow*, a loitering anti-radar munition.[500]

DoD Directive 3000.09 prohibits *lethal* fully autonomous robots. And semi-autonomous robots cannot "select and engage individual targets or specific target groups that have not been previously selected by an authorized human operator," even in the event that contact with the operator is cut off. Autonomous weapon systems may be used to apply non-lethal, non-kinetic force, such as some forms of electronic attack, against materiel targets; however, it specifically excludes:

> "Cyberspace systems for cyberspace operations; unarmed, unmanned platforms; unguided munitions; munitions manually guided by the operator (e.g., laser- or wire-guided munitions); mines; or unexploded explosive ordnance."[501]

[498] Andrea Gilli and Mauro Gilli, "The Diffusion of Drone Warfare? Industrial, Organizational and Infrastructural Constraints: Military Innovations and the Ecosystem Challenge," *Security Studies* 25, 2016: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425750.

[499] M. Hanlon, "Low-Cost Autonomous Attack System successfully flight tested," *New Atlas,* 4 Nov 2005.

[500] C. Kopp, "Precision guided munitions: Rockwell AGM-130A/B and Northrop AGM-136A Tacit Rainbow, *Air Power Australia*, May 1988.

[501] DoDD 3000.09, pp. 1-2. It has been pointed out that this seemingly well-defined policy distinction of applicability may nonetheless introduce a disconnect into DoD policy with respect to "unarmed, unmanned platforms," since such systems, if they malfunction, maystill inflict injury or collateral damage to individuals and property. For example, a malfunctioning

Levels of autonomy

Various definitions, classification systems, models, performance assessment metrics, and taxonomies of autonomy have been proposed over the last several decades, each with their own advantages, disadvantages and potential applications to military operations. The first categorization scheme was introduced by Sheridan and Verplank[502] in 1978 (and subsequently expanded by Sheridan[503] in 1992), and still used as the basis of many modern incarnations—organizes autonomy according to a 10-point scale, with higher numbers denoting higher levels of autonomy (e.g., on level 10, the machine acts completely on its own), and lower levels denoting decreased autonomy (e.g., on level 1, the human operator has full control):

- *Level 1* = computer offers no assistance; the human must make all decisions and take all actions
- *Level 2* = computer offers a complete set of decision/action alternatives
- *Level 3* = Level 2 + narrows the selection down to a few
- *Level 4* = Level 2 + suggests one alternative
- *Level 5* = Level 4 + executes that suggestion if the human operator approves
- *Level 6* = Level 4 + allows the human a restricted time to veto before automatic execution, or
- Level 7 = Level 4 + allows the human a restricted time to veto before automatic execution, or
- *Level 8* = Level 4 + *informs* human after execution only if it is asked
- *Level 9* = Level 4 + informs human after execution only if it decides to
- *Level 10* = computer decides everything and acts fully autonomously, ignoring the human

automated convoy vehicle may injure a person or cause damage that is similar in its effect to collateral damage from an errant autonomous weapon system. This is but one instance of a slew of ambiguity-ridden and ethics-related issues regarding the use of autonomy, a few of which are discussed later in this report. Ref:J. Caton, *Autonomous Weapon Systems: A Brief Survey of Developmental, Operational, Legal, and Ethical Issues*, U.S. Army War College, Strategic Studies Institute, Carlisle, PA, Dec 2015.

[502] T. Sheridan and W. Verplank, *Human and Computer Control of Undersea Teleoperators*, Man-Machine Systems Laboratory Report, MIT, 1978.

[503] T. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, MIT Press, 2003.

An alternative to this basic construct (introduced by Ensley and Kaber[504] in 1999) uses a similar 10-level scale, but is explicitly organized according to four basic functions: (1) *monitoring* (i.e., scanning displays); (2) *generating* (i.e., formulating options or strategies to meet goals); (3) selecting (i.e., adjudicating an option or strategy); and (4) *implementing*, or acting upon the selected option. A more recent variant, proposed by Parasuraman, et al.,[505] proceeds from the observation that previous taxonomies focus too much on a system's output functions (i.e., to decision and action) at the expense of input functions (i.e., sensing and gathering information).

Though also including an autonomy scale that varies from low to high, unlike prior categorizations, Parasuraman, et al. apply their scale to specific types of functions (that describe different stages of automation): (1) *acquisition* (of information), (2) *analysis*, (3) *decision and action selection*, and (4) *implementation* (of action). The acquisition stage includes systems that scan and observe the environment; analysis includes tasks similar to what human operators would naturally do (e.g., summarizing and fusing disparate pieces of information, predicting future state of environment, and other manipulations of gathered data); decision refers to adjudicating among possible courses of action (e.g., choosing a navigational route from several alternatives); and action implementation refers to automation that actually executes the selected action(s). Parasuraman, et al.'s taxonomy adds primary and secondary evaluative criteria, designed to evaluate the *need* for automation in the context of what can be generally expected of a human operator. Primary criteria consist of measures of consequences of human performance (e.g., mental workload, situation awareness, complacency, etc.); secondary criteria focus on consequences of automation (e.g., reliability, costs, etc.). In practice, the taxonomy has to be applied iteratively: first the primary criteria are evaluated, and the level of automation is adjusted accordingly; next, the secondary criteria are evaluated, and the level of automation is adjusted again. The goal was to define an objective method to determine a level of automation that is appropriate for a specific system.

Of course, myriad other generalizations of these basic multi-level schemes exist.[506] Some emphasize the details related to a system's output functions (i.e., to its decision capability); others focus on making detailed distinctions among input

[504] M. Endsley and D. Kaber, "Level of automation effects on performance, situation awareness and workload in a dynamic control task," *Ergonomics* 42, no. 3, 1999.

[505] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. on Systems Man and Cybernetics Part A: Systems and Humans* 30, no. 3, 2000.

[506] J. Beer, A. Fisk, and W. Rogers, "Toward a Psychological Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction* 3, 2014.

functions, such as how the system acquires information and formulates options. Most variants of this basic multi-level construct lack specificity for intermediate levels of autonomy, and generally suffer from having too many subjective components.

On the one hand, autonomy is far from being a "new" concept. Autonomy has been part of military systems for over three quarters of a century, with the first guided munitions appearing during World War II, and human-supervised automated defensive systems now being used by many militaries throughout the world.[507] On the other hand, as more and more AI-derived technologies find their way into modern weapon systems, and as the complexity, decision-making, and problem-solving capabilities of machines (which, as discussed earlier, now frequently exceeds human performance in specific problem domains) continue to increase, autonomy as a "single-word" concept is myopically shallow, at best, and misleading, at worst. Even if used as a placeholder for split-level gradations in meaning, its utility is limited since all such gradations involve subjective elements, and none (as of this writing) include a mission-centric context. Indeed, the U.S. Defense Science Board's 2012 report on autonomy suggests doing away with defining "levels" of autonomy altogether, and to instead shift the focus from "viewing autonomy as an intrinsic property of unmanned systems in isolation, [to] the design and operation of unmanned systems … in terms of human-systems collaboration."[508] The report recommends replacing definitions and levels with a conceptual framework (see next section).

Categorizations of autonomy based purely on innate autonomous *behaviors* (that is, conceptualized in terms of what an autonomous system does, independent of context) include those based on: (1) "sense, plan, and act" (SP&A) primitives, (2) "think, look, talk, move, and work" (TLTM&W) primitives, and (3) John Boyd's "observe, orient, decide, and act" (OODA) loop concept.[509]

SP&A models were popular mostly during the 1960s through 1980s, but were eventually supplanted by a behavior-based robotics approach, which is characterized by low level sensor-action loops (i.e., the "plan" primitive plays no role).[510] Beer, et al.,

[507] B. Watts, *Six Decades of Guided Munitions and Battle Networks: Progress and Prospects*, Center for Strategic and Budgetary Assessments, March 2007.

[508] *The Role of Autonomy in DoD Systems*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

[509] Col. John Boyd, USAF, *The Essence of Winning and Losing*, Briefing Slides, 28 June 1995: https://web.archive.org/web/20110324054054/http://www.danford.net/boyd/essence.htm.

[510] R. Arkin, *Behavior-Based Robotics*, MIT Press, 1998.

have recently introduced a 10-level autonomy scale based on the SP&A primitives (see figure 27).

Figure 27. Sense-Plan-Act-based levels-of-autonomy (H = human, R = robot)*

Level	Sense	Plan	Act	Description
1 - Manual	H	H	H	The human performs all aspects of the task including sensing the environment, generating plans/options/goals, and implementing processes
2 - Tele-operation	H/R	H	H/R	The robot assists the human with action implementation. However, sensing and planning is allocated to the human. For example, a human may teleoperate a robot, but the human may choose to prompt the robot to assist with some aspects of a task (e.g., gripping objects).
3 - Assisted Teleoperation	H/R	H	H/R	The human assists with all aspects of the task. However, the robot senses the environment and chooses to intervene with task. For example, if the user navigates the robot too close to an obstacle, the robot will automatically steer to avoid collision
4 - Batch Processing	H/R	H	R	Both the human and robot monitor and sense the environment. The human, however, determines the goals and plans of the task. The robot then implements the task
5 - Decision Support	H/R	H/R	R	Both the human and robot sense the environment and generate a task plan. However, the human chooses the task plan and commands the robot to implement actions
6 - Shared Control With Human Initiative	H/R	H/R	R	The robot autonomously senses the environment, develops plans and goals, and implements actions. However, the human monitors the robot's progress and may intervene and influence the robot with new goals and plans if the robot is having difficulty
7 - Shared Control With Robot Initiative	H/R	H/R	R	The robot performs all aspects of the task (sense, plan, act). If the robot encounters difficulty, it can prompt the human for assistance in setting new goals and plans
8 - Executive Control	R	H/R	R	The human may give an abstract high-level goal (e.g., navigate in environment to a specified location). The robot autonomously senses environment, sets the plan, and implements action
9 - Supervisory Control	H/R	R	R	The robot performs all aspects of task, but the human continuously monitors the robot, environment, and task. The human has override capability and may set a new goal and plan. In this case, the autonomy would shift to executive control, shared control, or decision support
10 - Full Autonomy	R	R	R	The robot performs all aspects of a task autonomously without human intervention with sensing, planning, or implementing action

* J. Beer, et al., "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction*, Vol. 3, No. 2, 2014.

The most recent TLTM&W model was introduced by the Army Research Laboratory's (ARL's) Robotics Collaborative Technology Alliance (RCTA) consortium,[511] and is summarized in figure 28 (which also includes some technical challenges associated with each of the five primitives.

[511] *Robotics Collaborative Technology Alliance (RCTA): 2012 Annual Program Plan*, Army Research Lab, March 2012: https://www.arl.army.mil/www/pages/392/RCTA_FY12_APP.pdf.

Figure 28. Think-Look-Talk-Move-Work-based levels-of-autonomy*

	Enhanced Capability	Vision	Challenges
"Think"	*Adaptive Tactical Reasoning*	Robots understand the concept of a mission or task, including stages of progress and measures of success	Adaptive tactical reasoning requires both declarative and procedural knowledge with which to reason. Neither exists in current systems, which generally have no data structures for mission level information. Tactical reasoning also requires some kind of model of the other members of the team, both human and robot, so that reasonable predictions of expected behavior can be made.
"Look"	*Focused Situational Awareness*	Autonomous ground systems maintain SA that is relevant to the current task and the larger mission	Focused SA, requires a semantic/cognitive description of the robot's environment that current systems do not have. SA also requires a sense of salience, what is important based on a shared understanding among teammates. Better learning is needed to develop a more human-like hierarchical understanding of object categories in the first place as well as to refine perception capabilities in the field.
"Talk"	*Efficient Proactive Interaction with Humans*	Robots interact with each other and especially with soldiers in an efficient and proactive way relevant to the evolving situation	Existing robotic systems are notoriously opaque and distrusted. They cannot explain what they are doing, primarily because they do not have meta-cognition; in other words, they do not have a model of their own behavior. Current systems also lack the ability to understand human (i.e., semantic) communication of orders or other information.
"Move"	*Safe, Secure, Adaptive Movement*	Robots that move on orders or their own initiative from one tactical position to the next with little or no reliance on metric inputs such as GPS	Current systems have insufficient descriptions, or models, of the world in which the robot is moving. Useful movement is also hampered by the lack of task or mission context so that a robot may persist in trying to reach a particular location that is not needed for the mission. Robots also need to be able to move in crowded and unpredictable environments, where existing algorithmic approaches are probably intractable but new learning approaches may work.
"Work"	*Interaction with the Physical World*	Robots are able to observe objects at close quarters to enable 3D interaction with them. They pick-up and move objects, either upon semantic direction or their own initiative.	The top four capabilities (*think-look-move-talk*) largely enable the performance of the main goal of the mission – the "work" the robot is to do. The work most often involves direct physical interaction with the world: entering and searching a building or vehicle, loading and delivering supplies, inspecting a suspected IED, etc. There is generally great uncertainty about the objects with which the robot is attempting to interact.

* Robotics Collaborative Technology Alliance (RCTA): 2012 Annual Program Plan, Army Research Lab, March 2012: https://www.arl.army.mil/www/pages/392/RCTA_FY12_APP.pdf.

Figure 29 shows a basic *Observe-Orient-Decide-Act* (OODA) loop overlaid with elements pertaining to properties expected of autonomous systems. The OODA loop is a simple model of decision making introduced by USAF Col. John Boyd in the late 1960s.[512] Intended originally as a conceptual backdrop to facilitate discussion and analyses of air combat (and military strategy in general),[513] it has since been applied to widely diverse fields that involve decision-making in adversarial environments (e.g., business, law enforcement, and sports).[514] The OODA loop has also been used to model planning and human supervisory control of physical systems.[515]

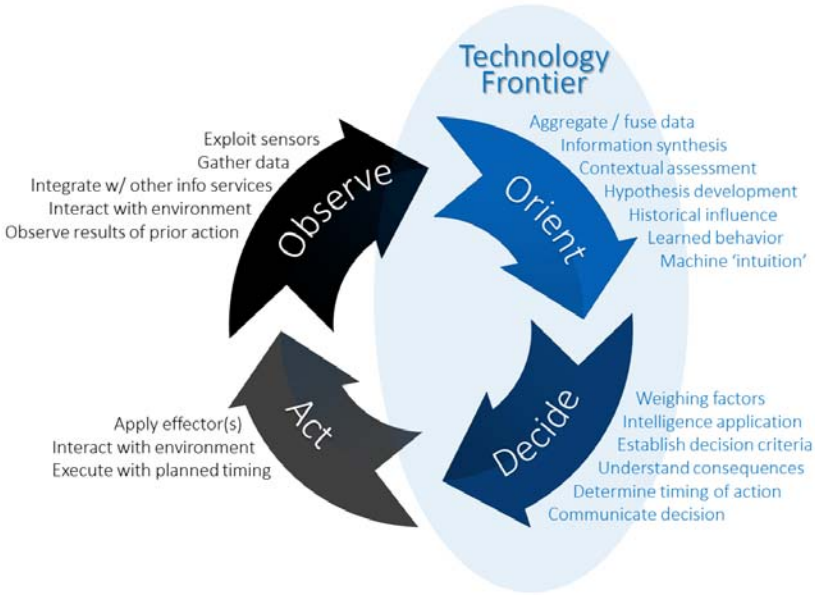
[512] R. Coram, *Boyd: The Fighter Pilot Who Changed the Art of War*, Back Bay Books, 2004.

[513] The OODA loop was originally used to help understand why American fighter pilots were more successful than their adversaries in the Korean War. Although MiG-15's were technically superior to the American F-86 Sabres, Boyd argued that it was because of the F-86's superior cockpit visibility that American pilots were able to decide and act faster than their opponents; i.e., American pilots were generally able to get into a good firing position before their Korean countgerparts could react. Ref: F. Osinga, *Science, Strategy and War*, Routledge, 2006.

[514] G. Hammond, *The Mind of War: John Boyd and American Security*, Smithsonian Books, 2004.

[515] T. Grant, "Unifying Planning and Control using an OODA-based Architecture," *Proceedings of SAICSIT*, 2005

Figure 29. OODA loop and elements pertaining to properties expected of autonomous systems*



* After figure 1 in M. Francis, "Unmanned air systems: challenge and opportunity," 2011 AIAA Wright Brothers Lecture, *Journal of Aircraft*, Vol. 49, No. 6, Nov-Dec 2012.

Observe refers to the stage during which information about the environment is collected, including characteristics of the physical environment, and the disposition, capabilities and intentions of enemy, friendly, and noncombatant forces; the *Orient* stage consists of aggregating, correlating and analyzing collected information and compiling a real-time situational awareness picture; the *Decide* step involves weighing various factors and options against the assigned (and/or locally determined) objectives to determine a course of action; and the *Act* step consists of following through on the decision (e.g., striking a target, applying navigational course correction, or engaging radar jamming). Implicit in the OODA construct is that it is not a static loop, but rather consists of multiple interlinking processes containing many inputs and decision steps. It is a continuous loop, with myriad simultaneous decisions that must be adjudicated in parallel. Information, too, is a fluid construct that is transformed continuously throughout all parts of the cycle, and by both sides of an adversarial confrontation. Indeed, one of the fundamental challenges of information-based warfare is the adaptive control and management of data and decision processes distributed within a networked force.[516]

[516] R. Deakin, *Battlespace Technologies*, Artech House, 2010.

An autonomous system must perform an analogous set of functions to what a human needs to accomplish a given task. While the *Observe* and *Act* phases are straightforward robotic "stand ins" for their human-centric counterparts (mechanical sensors act as surrogates for human perception, and actions are executed by one or more robotic effectors)[517] the functions that make up the *Orient* and *Decide* phases of the OODA-loop—highlighted in blue in figure 29—contain the autonomous system's key AI and other computation-based capabilities.

Orient is arguably the richer of the two (in terms of the complexity of the required algorithms), since it involves functions that determine how well the system is able to "understand" its environment: aggregating and fusing asynchronous data from multiple data sources; hypothesizing about, and deducing features to describe from current conditions; and incorporating historical data, past and/or learned experience in "making sense" of a situation. Higher levels of autonomy may require systems to make reasoned inferences and abductions (recall earlier discussion on page []), something that state-of-the-art algorithms are getting increasingly better at. General "perception" algorithms that are able to fuse and draw inferences from multiple forms of sensory input at the edge of what has been achieved as of this writing.

The *Decide* stage requires almost as complex a set of functions as *Orient*, since it includes such tasks as choosing (and applying) an appropriate set of features and weights to accommodate real-time decision making; adapting decision criteria to a dynamic environment (that may contain elements that themselves evolve according to disparate time-scales; and (borrowing from *DeepBlue*'s "look ahead" capability in chess) anticipating adversarial countermeasures in the next *Decision* cycle. Many of these elements are, just for those just recounted for the *Orient* stage, at the cutting edge of AI capability. Specific methods (and requirements, depending on operational context) span the gamut from simple physics-based move-and-act rules to the most sophisticated AI-driven (and/or swarm-based) adaptive behaviors and real-time learning.

An 11-level autonomy taxonomy (introduced by the Air Force Research Lab) that is organized around the OODA loop is shown in figure 30.

[517] In robotics parlance, an "effector" is any device that affects the environment (e.g., legs, wheels, arms, fins); the details depend on the application of the robot. An "actuator" is the actual mechanism that enables the effector to execute an action (e.g., electric motors, hydraulic or pneumatic cylinders, etc.). "Effector" and "actuator" are sometimes (erroneously) used interchangeably to denote whatever mechanism is required to "make the robot take an action."

Figure 30. OODA-loop-based autonomy taxonomy*

Level	Description	Observe	Orient	Decide	Act
		Perception / Situational Awareness	Analysis / Coordination	Decision Making	Capability
0	Remotely piloted vehicle	Flight control (attitude, rates) sensing; on-board camera	Telemetered data; remote pilot command	N/A; off-board pilot	Control by remote pilot
1	Execute preplanned mission	Preloaded mission data; flight control and navigation sensing	Pre/post flight BIT; report status	Preprogrammed mission and abort plans	Wide airspace separation requirements
2	Changeable mission	Health/status sensors	RT health diagnosis (Does UAV have problem?); off-board replanning (as required)	Execute preprogrammed or uploaded plans in response to mission and health conditions	Self accomplishment of tactical plan as externally assigned
3	Response to real-time faults/events	Health/status history and models	Tactical plan assigned; RT health dialog; compensate for most control failures and flight conditions	Evaluate status vs. required mission capabilities; abort/return to base if insufficient	Self-accomplishment of tactical plan as externally assigned
4	Fault/event adaptive vehicle	Off-board awareness – friendly system communicate data	All below plus ROE assigned; inner loop changes reflected in outer loop performance	On-board trajectory replanning – event driven; self resource management; deconfliction	Self-accomplishment of tactical plan as externally assigned
5	Real-time multi-vehicle coordination	Sensed awareness – local sensors to detect external targets (friendly and threat) fused with off-board data	All below with prognostic health management; group diagnosis and resource management	On-board trajectory replanning – optimizes for current and predictive conditions; collision avoidance	Group accomplishment of tactical plan – as externally assigned; air collision avoidance; possible close air space separation; formation in non-threat conditions
6	Real-time multi-vehicle coordination	Ranged awareness – on-board sensing for long range, supplemented by off-board data	All below plus enemy location sensed/estimated	Coordinated trajectory planning and execution to meet goals – group optimization	Group accomplishment of tactical goal with minimal supervisory assistance; possible close air space separation
7	Battlespace knowledge	Short track awareness – history and predictive battlespace data in limited range, timeframe, and numbers; limited inference supplemented by off-board data	Tactical group goals assigned; enemy location estimated	Individual task planning / execution to meet goals	Group accomplishment of tactical goal with minimal supervisory assistance
8	Battlespace single cognizance	Proximity inference – intent of self and others (friendly and threat); reduced dependence on off-board data	Strategic group goals assigned; threat tactics inferred; aided target recognition	Coordinated tactical group planning; individual task planning and execution; chooses targets of opportunity	Group executes mission with minimal supervisory assistance
9	Battlespace swarm cognizance	Knows intent of self and others (friendly and threat) in a complex/intense environment; on-board tracking	Group strategic missions assigned; threat tactics inferred	Distributed tactical group planning; individual mission decision-making; chooses targets	Group executes mission with minimal supervisory assistance
10	Fully autonomous	Cognizant of all within battlespace	Coordinates as necessary	Capable of total independence	Requires little guidance

* E. Sholes, "Evolution of a UAV Autonomy Classification Taxonomy," IEEE Aerospace Conference, 2007

A similar taxonomy based on the OODA loop (that uses 8 levels instead of the 10 levels shown in figure 30) was used by the National Aeronautics and Space Administration (NASA) as a method for helping assess an appropriate level of autonomy to design into a human spaceflight vehicle.[518]

[518] R. Proud, J. Hart, and R. Mrozinski, "Methods for Determining the Level of Autonomy to Design into a Human Spaceflight Vehicle: A Function Specific Approach," presented at *Performance Metrics for Intelligent Systems*, held 16-18 Sep 2003, Gaithersburg, MD.

Toward a conceptual framework of autonomy

As we have seen, there is no universally agreed upon definition of "autonomy." However, it may be possible to develop a general conceptual framework that can be used to both anchor theoretical discussions and serve as a frame-of-reference for understanding how theory, design, implementation, and operations are all interrelated. The framework would have to provide a method of objectively (or, as objectively as possible) distilling and convolving all of the individual key elements of autonomy.

To appreciate (at least on an intuitive level) how technically difficult a task it is to find an appropriate set of metrics to describe both *what* an autonomous system is and how *well* it is performing—not to mention the even more difficult task of developing a comprehensive conceptual framework in which these metrics have both a well-defined meaning for existing systems but are flexible and deep enough to anticipate the development of future systems—it is enough to recognize that the closer systems come to achieving full autonomy, the more closely aligned will any description of their behavior be to that of *describing the behavior of humans*. Therefore one ought not be surprised to learn that the autonomy-related research literature is replete with just about every combination of factors that may be used to categorize human, machine, and human-machine (hybrid) behaviors.

Different approaches may be distinguished according to which functions are emphasized at the expense of others. General categorizations include those that:[519]

- *Distinguish among the number of systems required for a given task* — in which a system's overall degree of autonomous functionality is tied to the number of vehicles required to complete a task; single-machine functionality is distinguished from multi-vehicle operations, from full swarm behavior.

- *Segregate functions according to the nature of the tasks involved* — in which tasks that are unique to the machine are distinguished from those that are machine-agnostic; machine intrinsic autonomous functions might include basic flight control, stabilization, and landing; machine-agnostic (i.e., task or mission oriented) functions might include navigation, route-planning, obstacle avoidance, and auto-determination of operational objectives.

[519] M. Francis, "Unmanned air systems: challenge and opportunity," *2011 AIAA Wright Brothers Lecture, Journal of Aircraft* 49, no. 6, Nov-Dec 2012.

- *Emphasize the degree of objectivity required in the decision-making process* — in which deterministic physics-based autonomous functions (e.g., stabilization, takeoff, landing) are distinguished from functions that must accommodate uncertainty and stochastic elements (variability in weather, turbulence, and other unpredictable elements, including enemy actions).

- *Evaluate functionality in the context of situational difficulty* — in which an autonomous system's ability to perform in complex but "simply characterized" environments (the objective description of which entails little or no uncertainty, and for which physics-based methods are appropriate) are distinguished from its ability to perform in highly uncertain environments (that require increasingly sophisticated AI).

- *Focus on the degree of complexity required of human-control* — in which the degree of autonomy is explicitly linked with the frequency (and sophistication) of command and control by the human operator.

Of course, none of these categorizations stands entirely apart from the others, and overlaps are unavoidable; e.g., the number of vehicles required to complete a mission is, in part, a function of environment, the degree of objectivity of decision-making is obviously dependent on context, and the complexity of human interaction arguably spans across all other categories. Indeed, with respect to this last category—and though it might be argued that any attempt to conceptualize autonomy for non-human systems must, inevitably, start with characterizing how non-human systems interact with human operators—even at this most basic level one finds disagreement in the literature:[520] some research groups start with the premise that higher robot autonomy entails lower levels (or less frequent) human-robot interaction (HRI); others assume that higher of system autonomy require higher levels (or more sophisticated forms) of HRI.[521] In reality, no objective one- or two-dimensional representation of autonomy can possibly capture the full extent of the functionality and behavior of autonomous systems. The DoD's Defense Science Board's (DSB's) 2012 report on autonomy (DSB/2012) argues that while a "level of autonomy"

[520] J. Beer, A. Fisk, and W. Rogers, "Toward a Psychological Framework for Levels of Robot Autonomy in Human-Robot Interaction," *Journal of Human-Robot Interaction* 3, 2014.

[521] Autonomy frameworks based on the second viewpoint also exist, but, because they are typically focused more on establishing guidelines for robots that assist humans in social situations, and are therefore less relevant for military operational contexts, will not be considered in this paper. Ref: S. Thrun, "Toward a framework for human-robot interaction," *Human-Computer Interaction* 19, 2004; M. Goodrich and A. Schultz, "Human-robot interaction: A survey," *Foundations and Trends in Human-Computer Interaction* 1, no. 3, 2007; R. Murphy and D. Woods, "Beyond Asimov: The Three Laws of Responsible Robotics," *IEEE Intelligent Systems*, July-August, 2009.

description may capture the essence of what makes a system autonomous, it fails to describe the specific milestones necessary to design and build such systems; moreover, it "deflects focus from the fact that all autonomous systems are joint human-machine cognitive systems, thus resulting in brittle designs." [522] In the end, DSB/2012 recommends that the DoD should abandon the debate over definitions of levels of autonomy altogether, and instead focus on developing a conceptual framework to help program managers and acquisition officers (and developers) holistically shape technology programs.

What all of the above categorizations have in common is that they all involve aspects of the "complexity" (objective measures of which are "to be determined" of course!) of various components of the coupled human-machine-mission-environment space:

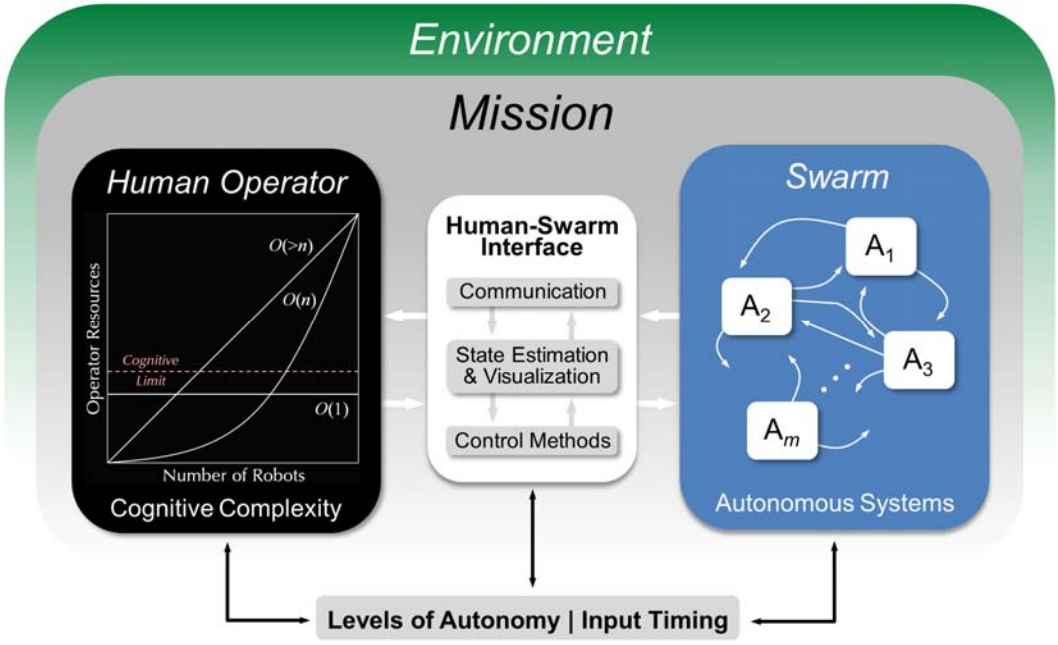
- The complexity of an autonomous system as a *physical machine*
- The complexity of autonomous swarm as a *system-of-systems*
- The complexity of the *environment* the autonomous system interacts with and makes adaptive decisions in
- The complexity of the *decision space* the system's AI-logic must contend with
- The complexity of the human↔machine *command-and-control relationship*

Figure 31 emphasizes the importance of the mission space and operational environment by embedding an earlier schematic depiction of the human-machine "system" (see figure 23) within a broader context, details of which appear in the ensuing discussion.

We conclude this section by introducing two most recent—and, as of this writing, *only* extant—attempts at formulating a multidimensional conceptual framework for assessing autonomy in unmanned systems: ALFUS and ASRF.

[522] *The Role of Autonomy in DoD Systems*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

Figure 31. Key components of human-swarm behavior and control (figure 23) embedded within broader context of environment and mission



After A. Kolling, et al., "Human Interaction with Robot Swarms," *IEEE Transactions on Human-Machine Systems*, Vol. 46, Feb. 2016.

ALFUS

The ALFUS (Autonomy Levels for Unmanned Systems) framework was developed at the National Institute of Standards and Technology (NIST) to serve as a framework to facilitate characterizing and articulating autonomy for unmanned systems.[523] ALFUS offers standard terms and definitions for requirements analysis and specification, and contains tools for evaluating and measuring the performance of unmanned autonomous systems. The framework is not a specific test or set of metrics, rather it represents one model of how a set of multidimensional metrics can be combined to generate an autonomy level. ALFUS's definitions are kept deliberately broad to encompass a variety of specific domains, ranging from logistics, manufacturing,

[523] H. Huang, E. Messina, J. Albus, *Autonomy Levels for Unmanned Systems (ALFUS) Framework,* Volume II: Framework Models Version 1.0, Ad Hoc Autonomy Levels for Unmanned Systems Working Group, National Institute of Standards and Technology, 2008: http://www.nist.gov/customcf/get_pdf.cfm?pub_id=823618.

search and rescue, medicine, and military applications. Introduced at the 2004 International Society for Optics and Photonics (SPIE) Defense and Security Symposium,[524] ALFUS is still under active development (as of this writing), though has not had significant changes over the last few years.

The general framework includes the following four components: [525]

1. Basic terms and definitions
2. Detailed model for autonomy levels
3. Summary model for autonomy levels
4. Guidelines, processes and use cases

ALFUS's detailed model derives from a three-dimensional decomposition of autonomous capability called Contextual Autonomous Capability (CAC); see figure 32.

Each axis refers to a particular metric group: (1) *mission complexity* (MC), (2) *environmental complexity* (EC), and (3) (the complexity of the) *human interface* (HI; or human independence). The decomposition is interpreted on two layers of abstraction:[526] *low level*, on which a system is characterized by a set of metric scores, such as the percentage of a mission that is planned and executed by the UMS onboard processors, the levels of task decomposition, and how easy it is to find a solution in the operating environment; and *above-metric level*, on which a system is assigned a total autonomy "score," which is effectively a weighted average of the individual metric scores.

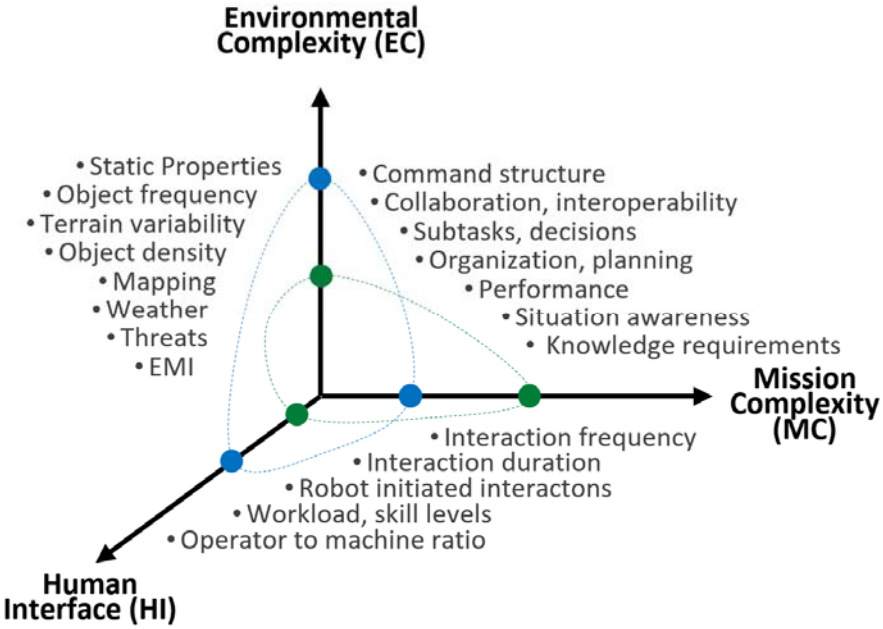
ALFUS flexibly provides additional layers of details below the metrics layer to facilitate application to specific systems. For example, the metric of human interface (HI) may be decomposed to actuation time, monitoring time, sensory data acquisition time, etc. Generally, the higher layers facilitate requirements specification and communication requirements, while the lower levels facilitate implementation and testing and evaluation.

[524] H. Huang, et al., "Specifying Autonomy Levels for Unmanned Systems: Interim Report," *Proceedings of the 2004 SPIE Defense and Security Symposium*, Orlando, Florida, 2004.

[525] H. Huang, *Autonomy Levels for Unmanned Systems (ALFUS) Framework,* Volume I: Terminology, Version 2.0, Ad Hoc Autonomy Levels for Unmanned Systems Working Group, National Institute of Standards and Technology, 2008: https://www.nist.gov/document-8274.

[526] Ibid.

Figure 32. Schematic of ALFUS's Contextual Autonomous Capability (CAC)



Ref: figure 5 in http://www.nist.gov/customcf/get_pdf.cfm?pub_id=823618

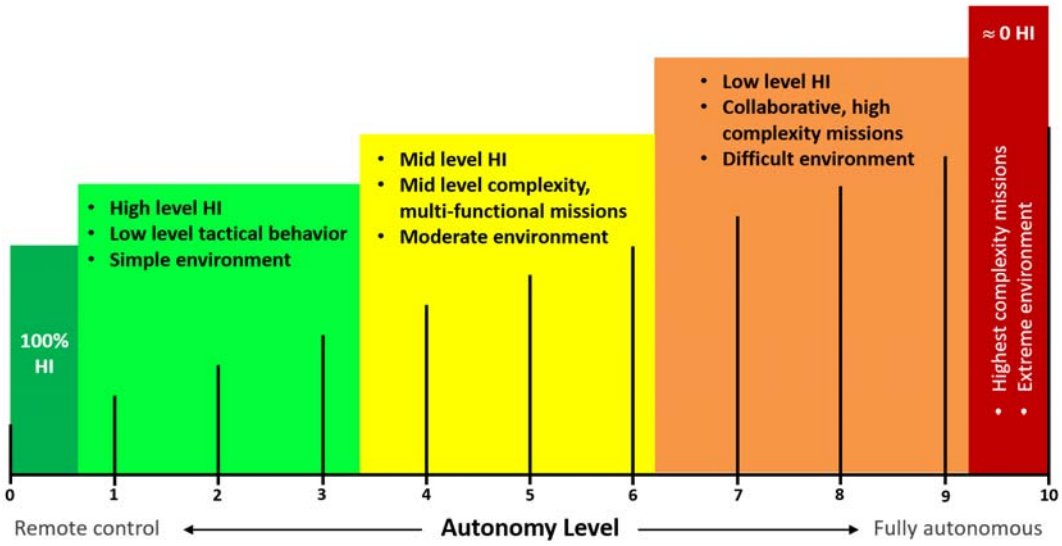
Note that CAC is devoid of any real meaning (certainly is incapable of assigning any measures of autonomy by itself) until one determines: (1) *what makes a mission complex*, (2) *what makes an environment complex*, and (3) *what makes an unmanned system human-independent*. Different sets of factors and (by implication) different sets of associated weights will result in different ALFUS-facilitated CACs, and, in turn, different assessments of a given system's level of autonomy.[527] For example, factors that may contribute to mission complexity include: mission time, required level of collaboration, synchronization of events, rules of engagement, sensory and data processing requirements, etc.; factors that may make an environment complex include the signal environment (including electromagnetic interference), the dynamic nature of the environment (including stigmergy), meteorological conditions, light, terrain, etc.; and factors that may contribute to making an unmanned system human-independent include increasing ability to sense wider areas of the environment, increasing ability to understand and analyze dynamic contexts, increasing ability to

[527] H. Huang, E. Messina, J. Albus, *Autonomy Levels for Unmanned Systems (ALFUS) Framework,* Volume II: Framework Models Version 1.0, Ad Hoc Autonomy Levels for Unmanned Systems Working Group, National Institute of Standards and Technology, 2008.

generate high-level complex plans, and an increasing ability to communicate with other systems without oversight.

Figure 33 shows a schematic distillation of ALFUS' three-dimensional CAC decomposition into a single-value of autonomy. The highest level of autonomy is one in which the system "completes all assigned missions with highest complexity; understands, adapts to, and maximizes benefit/value/efficiency while minimizing costs/risks on the broadest scope environmental and operational changes; and is capable of total independence from operator intervention."[528]

Figure 33. Schematic distillation of ALFUS' three-dimensional CAC decomposition into a single-value of autonomy



HRI = Human-robot interface. Ref: figure 4 in http://www.nist.gov/customcf/get_pdf.cfm?pub_id=823618.

A Mid-level autonomous system can plan and execute tasks to complete an operator defined mission, has a limited understanding of and response to environmental information and operational changes, and relies on about 50% operator input. The lowest-level autonomous system requires remote control even for simple tasks in simple environments. Note that the autonomy level generally refers to the HI aspect, with the other two axes serving as context. ALFUS's design allows for autonomy level

[528] H. Huang, K. Oavek, J. Albus, and E. Messina, "Autonomy Levels for Unmanned Systems (ALFUS) Framework: An Update," *SPIE Defense and Security Symposium*, Orlando, Florida, 2005.

to be interpreted as a nominal value, while instantaneous values may be dynamic, as a system adapts to specific mission conditions and changes in the environment. No attempt is made to describe the detailed differentiation among consecutive levels.

Of course, none of these levels, or interpretations are well-defined. As of this writing, only a few standard bench tests exist to facilitate filling in even the basic CAC axes; and there are, as yet, no analytical methods (at least none that have been incorporated into ALFUS; other stand-alone attempts exist and are referenced later) for objectively combining the orthogonal metrics into a single-valued measure of "autonomy." Indeed, ALFUS utility derives principally from its *generality* as a conceptual framework. In its current form, it is little more than a skeletal structure that begs further brainstorming and elucidation. However—anticipating a discussion of how DoD's existing acquisition process does not easily accommodate the lifecycle of autonomous systems (that appears a bit later in this report)—if the approach described here is generalized to accommodate specific mission operational requirements (across the individual military Services), an ALFUS-like framework can be used in concert with (and to inform) an autonomous-system-centric generalization of DoD's acquisition process.

Advantages of the ALFUS approach include:[529]

- It does not myopically focus just on a system's innate physical characteristics
- It includes a notional "mission space" as a context for assigning autonomy
- It is not tied to a specific domain (leaving many options for generalizing the basic method to, say, "realistic" mission types and objectives)
- It represents a sustained, long-term effort to define a common standard for autonomy (albeit one that both demonstrates the *possibility* of such an endeavor, and the difficulty of the overall task, given that ALFUS has been under development for well over a decade).

Shortcomings of the ALFUS approach include:[530]

- Since ALFUS is an ongoing effort (albeit one whose details have not changed much over the last five years or so), all results are provisional
- ALFUS does not provide any tools to decompose tasks in a standard way

[529] E. Sholes, "Evolution of a UAV Autonomy Classification Taxonomy," presented at IEEE Aerospace Conference, IEEE, 2007

[530] P. Durst and W. Gray, *Levels of Autonomy and Autonomous System Performance Assessment for Intelligent Unmanned Systems*, US Army Corps of Engineers, Engineer Research and Development Center, Geotechnical and Structures Laboratory, ERDC/GSL SR-14-1, April 2014.

- While ALFUS provides a general method to assess the autonomy of a system, it does not provide an objective way to map a system's autonomous capabilities onto an overall autonomy level (i.e., there remain irreducibly subjective components to the decision process)
- ALFUS may provide *too much* granularity (offering, in principle, $10^3 = 1000$ different possible levels of autonomy for a given system); there is no provision for any domain-specific scaling

Another basic shortcoming of ALFUS is that it does not account for the variation of *desirability* of a given level of autonomy with type of mission or environment. For example, if the mission objective is to detect and deactivate an explosive device, full autonomy may be less effective than a teleoperated robot. However, as currently configured, ALFUS simply assumes that higher levels of autonomy are de facto more desirable. In terms of its applicability to T&E issues, ALFUS remains "too vague and too complex… [and] does not provide any guidelines describing test procedures. The complexity that ALFUS adds to the T&E process by requiring the environment metrics to be measured hampers the ability to assess autonomy for a broad range of applications."[531]

Autonomous System Reference Framework

The Autonomous System Reference Framework (ASRF) is a notional conceptual framework for autonomous systems that was introduced in the DoD's Defense Science Board's (DSB's) 2012 report on autonomy (DSB/2012).[532] Intended only as a point of departure, it may also be regarded as an extension of the ALFUS framework (albeit one that takes only the top-most level view, and includes even less specificity than ALFUS).

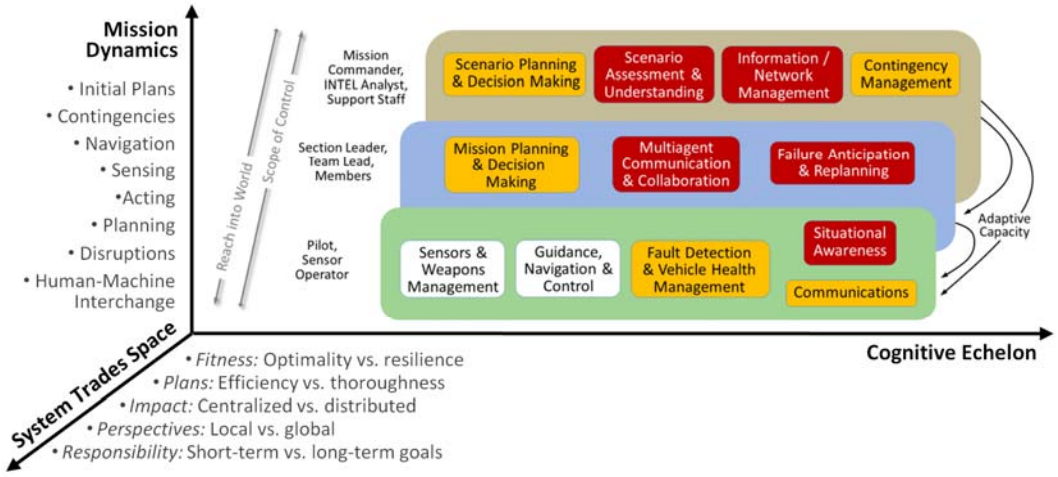
DSB/2012's point of departure is to visualize the challenges of autonomy from the perspective of three key stakeholders: the *commander*, the *operator*, and the *developer*. For the commander, the key issue is how a given mission will be accomplished, and for which autonomy matters only to the extent that it may alter how a mission must be managed. For the operator, autonomy is experienced directly, as in the human-machine interface and collaboration. And for the developer, autonomy is a euphemism for AI software (since that is where all autonomous behavior effectively originates); and, as such, generally falls outside DoD's current hardware-centric development and acquisition process.

[531] Ibid., p. 14.

[532] *The Role of Autonomy in DoD Systems*, DoD Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

A framework's ability to support the requirements specification, design, development, and review/approval phases of the acquisition process (discussed in a later section) is therefore contingent upon it embodying three classes of design decisions for autonomy that reflect the larger abstract space in which each of the three notional stakeholders live (see figure 34):

Figure 34. Schematic distillation of the Autonomous System Reference Framework



After section 3.3 and figures 3-1, 3-2, and 3-4 in DSB/2012.

1. The *cognitive echelon class* (CEC), which represents the space in which decisions on how autonomous systems are to be used are made (and includes different perspective of users with different spans of control; e.g., from vehicle pilot to section leader to mission commander).

2. The *mission dynamics class* (MDC), which represents the space in which the tradeoffs between potential benefits and challenges of autonomy may be described as functions of the times at which key decisions are made and specific actions executed (e.g., the degree of autonomy sufficient for take-off and landing may be different from what is required for navigation or targeting; and operators and systems may be generally expected to interchange initiatives and roles throughout a mission and across echelons, as the coupled human-machine system adapts to new events, disruptions and opportunities as they arise.

3. The *human-machine system trades space class* (HMSTSC), which denotes the space of factors that describe (and lie at the cusp between) human and autonomous system performance (e.g., reliability, manpower, and training costs).

DSB/2012's third class, HMSTSC, is introduced with a view toward providing developers and acquisition officers a tool for predicting unintended consequences (by linking potential symptoms of human-machine-interface imbalances), and describes it using a balloon metaphor: while autonomy (however it is defined) can increase the capability or capacity of a system, there are tradeoffs that can limit its expansion; i.e., that can "pop the balloon." Tradeoffs include:[533]

- *Fitness*: which describes the balance between optimal performance for expected missions and the need for resilience and adaptability for new missions and/or unexpected conditions; a possible "unintended consequence" is increased brittleness in the system

- *Plans*: which describes the balance between achieving an efficiency in following an existing plan with the ability to detect when a given plan is no longer valid (and being able to adapt); a possible "unintended consequence" is being locked into a wrong plan of action and/or increased difficulty in revising a plan

- *Impact*: which describes the balance between accessing and relying on viable information as a means toward achieving mission objectives and inadvertently putting a given mission at risk by incorporating nefarious information; a possible "unintended consequence" is a high cost of coordination

- *Perspectives* (i.e., ability to gain situation awareness): which describes the balance between focusing attention on local action on one system, from a human control point of view, with distributing and coordinating across multiple systems; possible "unintended consequences" include data overload and a reduced speed in decision making

- *Responsibility*, which describes the balance between short term and long term goals, and resolving between multiple-simultaneous goals (a perennial general technical challenge for AI-driven machine learning systems); an "unintended consequence" may be a break down in collaboration and coordination

The CEC is illustrated in figure 34 with elements that highlight the kinds of decisions made across echelons, and the varying levels of impact those decisions can have on a mission; e.g., increasing "scope of control" from bottom to top, and decreasing the effective "reach into the world" from top to bottom. There is also an implied adaptive capacity that reaches from upper to lower levels. For example, autonomy may use

[533] Ibid., Section 3.3.

higher-level routes and waypoints as references for controlling vehicles, and to translate raw sensor data into "higher levels" of information (e.g., targeting). At the highest levels, autonomy may be used to assist in allocating tasks, managing resources (e.g., assigning actions to individual sensors and/or defining weapon load outs), and processing and fusing raw Intelligence (INTEL) data. The point of the schematic shown in figure 34 (or of this discussion) is not to provide an exhaustive list of all possible ways in which autonomy can be used to assist the warfighter/analyst, but merely to provide a notional framework, much like ALFUS, that indicates where and when autonomy is—or *can* be—used.

The cognitive echelon component in figure 34 also highlights the relative status of technology deployment for a given capability: *orange* means that an existing capability is determined by DSB/2012 to be underutilized for a given mission component (e.g., scenario planning and decision making on the highest echelon, and vehicle health management on the pilot/sensor operator level), and *red* represents a an open technical challenge for which future investment is indicated.

DSB/2012 identifies six key areas in which advances in autonomous technology would have significant impact on the deployment and performance of unmanned systems:[534] (1) *perception* (including both hardware sensors and software-based "sensing"), (2) *planning*, (3) *learning*, (4) *human-robot interaction*, (5) *natural language understanding*, and (6) *multi-agent coordination*.

Perception

Perception for unmanned systems can be categorized into four (partly overlapping) classes: *navigation* (e.g., for guidance, navigation and control functions, to support path planning and dynamic replanning and to enable multi-agent communication and coordination), *mission sensing* (e.g., for mission planning, scenario planning, assessment and understanding, multi-agent communication and coordination and situational awareness), *system health* (e.g., for fault detection and vehicle health management,) and *manipulation* (which becomes increasingly important as missions move from perceiving-at-a-distance to acting-at-a-distance: e.g., opening doors, IED disposal, and general material handling).

Obvious advantages of increased autonomy for navigational perception include *vehicle safety* (since an autonomous system can react much faster in dangerous situations than humans), and a reduced cognitive workload of a human operator.

[534] Section 3.4 in *The Role of Autonomy in DoD Systems*, DoD DSB, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

Increased autonomy for mission sensing can potentially: (1) enable covert operations without network connectivity (reducing network vulnerability and the cognitive load of the human operator), (2) reduce the number of human analysts needed to assimilate large amounts of raw data (via autonomous recognition, cueing, and/or prioritization of areas of interest), (3) reduce network demands (using onboard identification, and/or prioritization of data), (4) enhance navigation (e.g., by enabling onboard adaptive "decisions" to be made to hover, loiter, etc.).

Increased autonomy for monitoring system health has at least three benefits: (1) enabling a more graceful degradation of performance than otherwise possible via human reaction), (2) increasing trust in the system (via self-monitoring for unexpected behaviors, particularly during critical phases of a mission), and (3) reducing the cognitive workload of operators (by freeing them from monitoring diagnostic displays).

Finally, increased autonomy for manipulation would both decrease the time and workload needed for manipulation tasks, and potentially reduce the number of robots needed for a mission (since a second robot would no longer be required to "observe" the dynamics between manipulator and the object being manipulated).

DSB/2012 identifies several critical gaps in current state-of-the-art autonomy technology for perception: (1) situation awareness in complex battle spaces (studies focus more on increasing navigational autonomy for individual or related swarms of platforms than on integrating battlespace-wide data, UGV navigation in urban environments, with dense foliage, and with people remains nascent, as does multi-sensor fusion and comprehensive world-modeling); (2) too strong a focus on the development of new sensors at the expense of developing enhanced algorithms for existing sensors, particularly vision (e.g., lack of high-speed obstacle detection and recognition in complex terrain); (3) evidential reasoning about system health (the state-of-the-art of which is adequate for detecting independent faults in systems for which a full model exists, but fails at detecting, identifying and recovering from multiple dependent faults; additional research is also needed to understand how less-than-complete models of a system can be made accurate enough to support evidential reasoning); and (4) perception for manipulation generally remains a major gap for UGV.

Planning

Planning is the process of deciding upon a course of action designed to achieve a desired mission objective (while minimizing resources). Planning also lies at the very heart of all multiagent-based (MBM) approaches to AI, and requires: (1) a mathematically precise representation of the factors and conditions that describe the environment (and that agents can "sense" and "react" to), and (2) algorithms that assign "weights" (that denote lesser or greater required attention) to elements of the

environment and an agent's internal state, to adaptively compute "optimal" resources and actions, subject to whatever limitations and constraints are deemed necessary.

While decades of basic MBM research has unearthed (and bequeathed) a veritable warehouse of insights and methodologies for AI-based planning, major gaps remain, and are likely to persist because of their fundamental nature. Specifically, there is no generally applicable technique for developing planning algorithms that account for multiple simultaneous objectives in dynamic environments (the nature of which cannot, In general, be accounted for a priori). Moreover, the objectives—and therefore weights, and action-adjudication algorithms—can (and typically will) also change as an autonomous system gains "experience" while operating in a not-fully-described environment. The technology that would allow for a graceful continuous remapping of an environment, and an "on the fly" adjustment to an autonomous system's onboard "reasoning faculty" is still nascent.

Learning

Machine learning (ML) lies at the heart of most state-of-the-art research into developing intelligent, autonomous systems (see discussion on pages []-[]). Indeed, some of the best-known AI "successes" in recent years (e.g., IBM's *Watson*[535] and Google's *AlphaGo*[536]) were made possible by advances in ML. However, a limitation that applies to *all* extant ML methods as they apply to "narrow AI" problems is that they are effectively "black boxes" that do not easily reveal the "logic" behind the "reasoning."[537] This may be innocuous when playing an AI-system in chess, say, but assumes an entirely new (and serious) dimension if the "narrow AI" in question is embedded within a military autonomous system. For example, how does one ensure (during, say, the testing and evaluation phase of DoD's acquisition process) that whatever autonomous system being developed will not perform "surprising" (i.e., unanticipated) actions during a mission?

A second issue—at least as egregious as displaying impenetrably surprising behaviors—is that otherwise well-performing "narrow AI" systems can also sometimes (and unpredictably) provide *bad* solutions to problems, with counter-

[535] A. Sostek, "Human champs of 'Jeopardy!' vs. Watson the IBM computer: a close match," *Pittsburgh Post Gazette*, 13 Feb 2011.

[536] S. Byford, "Google's AlphaGo AI beats Lee Se-Dol again to win Go series 4-1," *The Verge*, March 15, 2016.

[537] In the context of AI-based text-processing systems, MIT has recently introduced a method to train neural networks so that they provide rationales for their otherwise (and traditionally) opaque classifications. Ref: T. Lei, R. Barzilay, and T. Jaakkola, *Rationalizing Neural Predictions,* Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, 2016. https://people.csail.mit.edu/taolei/papers/emnlp16_rationale.pdf.

intuitive properties. For example, two recent studies of state-of-the-art visual classifiers show that: (1) changing an image that has already been correctly classified (say, that of a panda) in a way that is *imperceptible to humans* can cause a deep-learning neural network (DNNs) to classify the image as something entirely different (as, say, a gibbon),[538] and (conversely) (2) it is easy to produce images that are *completely unrecognizable to humans*, but that are "classified" by state-of-the-art DNNs with 99.99% confidence (e.g. labeling with certainty that white noise static is a lion).[539]

Finally, most current ML methods require vast amounts of supervised training data, the development of which requires significant time and effort of human subject matter experts (e.g., required to label and/or annotate large number of image and video exemplars for training). Unsupervised learning methods certainly exist—the general technique of *reinforcement learning*[540] and Harvard's recent method called *Turing Learning*[541] are but two examples—but their applicability to the kinds of unstructured dynamic environments that military autonomous systems are expected to perform well in remains uncertain.

Human-robot interaction

Human-robot interaction (HRI) is a still-nascent multidisciplinary field, and is a subset of a broader field of study of human-system interaction (where the "system" may consist of multiple simultaneous linked robots, including those both physically situated and virtual. DSB/2012 identifies six basic HRI research issues:[542] (1) how humans and robots communicate; (2) how to model the relationship between humans and robots; (3) how to study and enhance human-robot teamwork;[543] (4) how to

[538] C. Szegedy, et al., "Intriguing properties of neural networks," presented at the *International Conference on Learning Representations*, 2014: https://arxiv.org/abs/1312.6199.

[539] A. Nguyen, J. Yosinski, and J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *Computer Vision and Pattern Recognition* (CVPR), IEEE, 2015: http://www.evolvingai.org/files/DNNsEasilyFooled_cvpr15.pdf. The authors of this study believe that *all* AI techniques that derive from creating decision boundaries between classes (not just deep neural networks) are subject to this "self fooling" phenomenon.

[540] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[541] G. Templeton, "Turing Learning breakthrough: Computers can now learn from pure observation ," *ExtremeTech*, 30 Aug 2016.

[542] Section 3.4.4 in *The Role of Autonomy in DoD Systems*, DoD DSB, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

[543] Our earlier discussion of human control of robotic swarms is worth recalling in the present context. The literature on human supervisory control of multiple semi-autonomous robots is sparse, and much of it remains cutting edge. Just as there is currently no general method that

predict usability and reliability in the human-robot teaming; (5) how to capture and express the HRIs for a particular application domain; and (6) how to characterize end-users. The study of HRI thus spans a broad spectrum of research domains: unmanned systems, human factors, psychology, cognitive science, natural language processing, communication, computer supported work groups, and sociology. (The problem of developing a taxonomy and set of metrics to describe HRI is—as we will see later in this section—a microcosm of the problem of developing a framework for describing autonomy in general; a recent survey[544] references 29 papers containing no less than 42 metrics, most of which still fall short of providing measures for HRI-derived capabilities and not just capabilities that stem from robots alone.)

The key benefits to DoD on focusing on HRI and human-robot teaming, rather than on robots alone include: reduced cost of operating and designing platforms, increased adaptability to new situations, and (assuming successful teamwork can be established) faster performance of tasks with fewer errors. Two teamwork styles are possible: *remote telepresence* (in which a human operator works through the unmanned system to perceive and act in real-time at a distance) and *taskable agency* (in which the robot is delegated sole responsibility for the mission). "Trust" becomes a major factor for taskable agents (and is discussed briefly later, in the context of testing and evaluation of autonomous systems).

There are two key gaps in HRI for unmanned systems (emphasized by DSB/2012 to highlight the likelihood of increasing use of human-robot teaming): (1) *natural user interfaces*, to enable trusted human-system collaboration (e.g., operator control interfaces, interfaces that facilitate human-robot dialog, and visual-centric interfaces that make visible what the unmanned system is doing and why), and (2) *understandable autonomous system behaviors*. The latter denotes a fundamental limitation due to current state-of-the-art practice in the modeling, prescribing, predicting, and/or testing (the veracity of) the "behavioral space" of autonomous systems (aspects that relate to testing and evaluation are discussed later in this section).

maps rules that describe context-dependent actions of individual robots to desired group behaviors (see previous section), there are (as of this writing) no validated schemes for scalable, flexible, and adaptive human control of robot teams. Ref: K. Sycara, *Robust Human Interaction with Robotic Swarms*, presentation at ICAART 2016.

[544] R. Murphy and D. Schreckenghost, "Survey of metrics for human-robot interaction," presented at the *24th IEEE International Symposium on Robot and Human Interactive Communication*, 2015.

Natural language understanding

Natural language processing (NLP) is broad interdisciplinary field focused on developing methods by which humans and computers can communicate using conventional "human" languages. Dating back to the roots of computer science in the 1940s and 1950s,[545] today it is a mix of computer science, artificial intelligence, and computational linguistics.

Since natural language is obviously the preferred way for humans to interact with other humans, HRI researchers are naturally interested in finding ways for human operators to communicate—e.g., issue orders to, extract high-level information from, collaborate in real-time with—autonomous systems. The drawback is that meaningful communication entails *understanding* (on the part of the robot); and achieving understanding is difficult because of the inherent ambiguity of information in changing contexts. Traditionally, and because the state-of-the-art in natural language understanding (NLU) is adequate only for basic tasks (e.g., simple instruction codes using a limited vocabulary, such as been commercially available via Amazon's *Alexa* program, Apple's *Siri*, and Google's *Home*),[546] human-machine interfaces have relied primarily on graphical user interfaces (GUIs). However, it is easy to imagine situations (such as when a human operator's hands are engaged in some other concurrent activity) that a NLU-interface would be highly preferable.

NLP, by itself, does not denote any specific method or algorithm, and is instead best thought of as a label for a broad rubric of related techniques and research. Examples include (and ordered roughly by "degree of difficulty" of achieving): *text summarization*, in which a given document is distilled to manageably small summary; *named entity recognition*, which is the task of identifying text elements that belong to certain predefined categories, such as the names of persons, organizations, locations, expressions of times, etc.; *relationship extraction*, in which the relationship between various named parts of a chunk of text are identified ("an object O *belongs* to person P"); *semantic disambiguation*, in which a priori ambiguous meanings of words (or chunks of text) are automatically disambiguated from a deeper analysis of context and/or information that may be culled from an "ontology" (see discussion below); *sentiment analysis*, in which certain kinds of subjective information is extracted from a document or set of documents (e.g., extracting a range of emotional reactions to public events from social media posts); *speech*

[545] S. Lucci and D. Kopec, *AI in the 21st Century*, Mercury Learning and Information, 2013.

[546] S. Adee, "It's just common sense," *New Scientist* 232, no. 3094, 8 Oct 2016; J. Dunn, "We put Siri, Alexa, Google Assistant, and Cortana through a marathon of tests to see who's winning the virtual assistant race — here's what we found," *Business Insider*, 4 Nov 2016.

recognition, which refers to the textual representation of sound recordings of people speaking.

NLU, in which semantic content is extracted from free-form text and speech, falls towards the tail end of this ranked list, and is arguably "the" most difficult open-research problem of NLP. To highlight the enormous difficulty involved in developing NLU systems, recall two notable recent examples discussed in an earlier section: (1) IBM's *Watson*, and (2) CyCorp's *CyC*. *Watson* is an ongoing research effort (initiated in 2007 by IBM), whose original goal was to develop an AI system that performs sufficiently well on open-domain (free-form based) question-and-answering to compete with human champions at the game of *Jeopardy!*[547] In 2011, Watson beat the two highest ranked *Jeopardy!* players of all time in a two game match (played on 14/15 Feb 2011).[548] As a benchmark of the research and development time required to build a sophisticated question-and-answer system capable of defeating the best human players of *Jeopardy!*, it took IBM four years of dedicated effort by a staff of 20.[549] Although IBM's Watson research team is no longer focused on *Jeopardy!*, development work continues on applying the underlying learning method (deep learning and reinforcement learning)[550] to other problems (e.g., health care). The challenge has been to adapt Watson to a specific domain, which has proven to be a highly nontrivial enterprise (see earlier discussion).

The *Cyc* Project[551] (its name is derived from en–*cyc*–lopedia) is predicated on the idea that a HAL/9000-like artificial intelligence system[552] can only be developed by first codifying, in machine-usable form, a significant fraction of millions of pieces of knowledge that comprise human common sense (e.g., "an automobile is driven on a highway," "a playground is a place," and "a lemon is sour"). Started in 1984 by

[547] http://www.jeopardy.com/. Some sample *Jeopardy!* questions: (1) " On Sept. 1, 1715 Louis XIV died in this city, site of a fabulous palace he built" (ans: "What is Versailles?"); (2) "Pseudonym of labor activist & magazine namesake Mary Harris Jones" (ans: "What is Mother Jones?"); and (3) "Sakura cheese from Hokkaido is a soft cheese flavored with leaves from this fruit tree" (ans: "What is Cherry?"). Ref: "Sample 'Jeopardy!' questions," Arizona State Univ., 2010: https://asunews.asu.edu/20101103_jeopardyquestions.

[548] S. Baker, *Final Jeopardy: Man vs. Machine and the Quest to Know Everything*, Houghton Mifflin Harcourt, 2011.

[549] http://researcher.watson.ibm.com/researcher/view_group_pubs.php?grp=2099

[550] D. Ferrucci, et al., Building Watson: An Overview of the DeepQA Project, *AI Magazine*, 2010.

[551] http://aitopics.org/link/cyc-project.

[552] HAL/9000 is the fictional human-like artificial intelligence system depicted in the movie *2001: A Space Odyssey*.

Douglas Lenat at the Microelectronics and Computer Technology Corporation,[553] the *Cyc* Project moved under the auspices of an independent company, Cycorp, Inc. in 1994 (with Lenat still serving as President and CEO),[554] where it continues to do research on ontology, knowledge representation, knowledge acquisition, natural language processing and machine learning. Nearly half of Cycorp's revenue comes from U.S. government agencies.[555] An open source version of its knowledge base and functionality was released to the public under the OpenCyc project in 2002.[556] As of 2012, OpenCyc contains approximately 240,000 concepts and 2,000,000 "rule of thumb" assertions (the full *Cyc* knowledge Base contains around 500,000 concepts and 5,000,000 assertions). *Cyc* is currently the world's largest existing commonsense knowledge base ontology and symbolic-AI inference reasoning engine.[557]

A typical deductive inference by *Cyc* is "Bob is wet," generated from the statement "Bob is completing a marathon."[558] *Cyc* uses its commonsense rules to deduce, first, that a marathon is a form of race; second, that any human (i.e., "Bob") who runs a marathon must physically exert himself; third, that people sweat when exerting themselves; and, finally, that when anything sweats it is wet. However, *Cyc*'s ability to infer new knowledge from an existing base is limited. For example, the system does not currently support either *inductive* or *abductive* reasoning. Questions regarding the potential of *Cyc*'s long-term growth include: (1) the "brittleness" of its assertions (currently treated as simple binary true/false statements; i.e., there is no room for probabilistic "fuzziness"), and (2) the scalability (the system's performance must both degrade gracefully as the knowledge base grows large—*Cyc*'s current inference engine sometimes "slows down to a crawl" during particularly large searches through its entire knowledge base[559]—and be free of reification (i.e., the inadvertent interpretation of "possibly true" facts as definitely true). The "takeaway" is that though *CyC*, like *Watson*, arguably represents the current apex of the state-of-the-art,

[553] A history of the project is described by Lenat in an interview with S. Laningham: IBM DeveloperWorks, podcast, 16 September 2008: http://www.ibm.com/developerworks/ podcast/ dwi/cm-int091608txt.html.

[554] http://www.cyc.com/.

[555] L. Wood, "Cycorp: The Cost of Common Sense," *Technology Review*, March 2005.

[556] http://sourceforge.net/projects/opencyc/files/; http://www.opencyc.org/images/opencyc-kb-browser.gif.

[557] D. Ramachandran, P. Reagan, and K. Goolsbey, "First-Orderized ResearchCyc: Expressivity and Efficiency in a Common-Sense Ontology," in *AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications*, 2005.

[558] Example based on B. Copeland, *Cyc*, Encyclopedia Britannica online: http://www.britannica.com/EBchecked/topic/752898/CYC

[559] J. Friedman, "The Sole Contender for AI," *Harvard Science Review*, 2003

it is both limited in terms of its innate ability to "understand" and draw general inferences, and required a prodigious amount (25+ years) of research of development.

Multi-agent coordination

Multi-agent coordination refers to the problem of distributing a task over multiple autonomous robots, software agents, and/or humans. Group coordination may either emerge from the agents interacting or negotiating with each other either directly (distributed coordination) or by being explicitly directed by a human operator (centralized coordination). However coordination is achieved, the method must synchronize the activities of all agents involved, and accommodate real-time changes to the environment.

Apart from introducing potentially new CONOPS (such as swarm tactics), the ability to coordinate multiple autonomous systems has at least four benefits:[560] (1) *increased coverage*, (2) *decreased costs*, (3) *redundancy*, and (4) *specialization*. Multiple autonomous systems can provide persistent coverage over wider areas than individual platforms. Having many low-cost systems can potentially provide the same performance effectiveness as a single high-cost platform, while also providing redundancy to militate against dangers in contested areas. If multi-agent coordination can be combined with autonomous planning, each would enhance the benefit of the other (e.g., the actions of multiple robots could be coordinated in real-time in communications-denied areas).

The main caveat to achieving these capabilities was discussed in an earlier section in the context of state-of-the-art practices in engineering robotic swarms:[561] *no general method currently exists that maps individual rules to desired group behavior*. The design of multi-agent systems has, to date, been mostly ad-hoc, with little or no (meta) coordination of effort among different research groups (and specific algorithms and details of their implementation remaining proprietary).

Other gaps in technology include:

- *Unpredictability of emergent behavior*, which refers to one of the core attributes of complex adaptive systems (namely, that there is no general method to predict the global behavior of a system from knowing only its constituent parts and low-level rules of interaction)

[560] Section 3.4.6 in *The Role of Autonomy in DoD Systems*, DoD DSB, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

[561] I. Navarro and F. Matia, "An Introduction to Swarm Robotics," *International Scholarly Research Notes*, Vol. 2013, 2013: https://www.hindawi.com/ journals/isrn/2013/608164/.

- *Spectre of unanticipated interference*, refers to the possibility that individual robots will unintentionally interfere with one another (and is a partial consequence of the main caveat and gap #1)
- *Communication*, in the sense that the fundamental problem of identifying *what* to communicate (both to and within a robotic swarm), and *when* to communicate it, remains an open question in multi-agent design research
- *Levels of intelligence*, which refers to another open research problem that seeks to determine an "optimal" balance—which may need to be maintained dynamically, as a mission unfolds—between onboard intelligence (that enables individual robots to adapt to changing conditions), swarm intelligence (in which one member, or at most a few members, of a swarm are endowed with an ability to direct other members of the swarm), and external control (in which a human directs and coordinates behaviors).

Technical challenges

The academic, commercial, and military research communities have identified a number of outstanding technical challenges that must be solved (to varying degrees) in order to develop fully autonomous systems; most stem directly from the cadre of as-yet unsolved problems associated with AI and the behavior of complex systems:[562]

1. *Fundamental "Devil is in the details" R&D hurdles:* while it is easy to formally anoint an unmanned system as "autonomous" and engage in Einsteinian *Gedankan*[563] (or "Thought") experiments to examine, and develop CONOPS for, various operational environments—as though such systems already exist—in order to actually develop an autonomous system requires confronting many of the same fundamental problems that the academic and commercial AI and robotic research communities have struggled for decades to "solve." To survive and successfully perform missions, autonomous systems must be able to *sense*, *perceive*, *detect*, *identify*, *classify*, *plan*, *decide*, and *respond* to a diverse set of threats in complex and uncertain

[562] A. Finn and S. Scheding, *Developments and Challenges for Autonomous Unmanned Vehicles: A Compendium*, Springer-Verlag, 2010; J. M. Bradshaw, et al., "The Seven Deadly Myths of Autonomous Systems," *IEEE Intelligent Systems* 28, no. 3, 2013; and J. Bornstein, "DoD Autonomy Roadmap – Autonomy Community of Interest," Presentation at NDIA 16th Annual Science & Engineering Technology Conference, March 2015.

[563] S. Perkowitz, "Gedankenexperiment," *Encyclopedia Britannica Online*: https://www.britannica.com/topic/Gedankenexperiment.

environments (see below). While aspects of all of these "problems" have been solved to varying degrees, there is, as yet, no system that fully encompasses all of these features. The reason is simple: the anticipation of the perceived benefits of autonomy is not (yet) in sync with the reality of what is currently possible (or has yet been demonstrated) in the AI and robotics research communities.[564]

2. *Complex and uncertain environments:* autonomous systems must be able to operate in complex—possibly, a priori unknown—environments that possess a large number of potential states that can neither all be pre-specified nor be exhaustively examined or tested. Systems must be able to assimilate, respond, and adapt to dynamic conditions that were not considered during their design. This "scaling" problem—i.e., being able to design systems that are developed and tested in static and structured environments, and then have them perform as required in dynamic and unstructured environments—is highly nontrivial.

3. *Emergent behavior:* if an autonomous system is to be able to adapt to changing environmental condition, it must have a built-in capacity to learn; and to do so in an unsupervised fashion. Under such conditions, it may be difficult to predict, and be able to account for, the system's emergent behavior.

4. *Human-machine interactions:* the operational effectiveness of autonomous systems depends on the dynamic interplay between the human operator and the machine(s) in a given environment, and how the system responds, in real-time, to changing operational objectives as the human adapts to dynamic contexts. The innate unpredictability of the human component in human-machine collaborative performance only exacerbates the other challenges identified by this list.

5. *Human-machine communication*: the interface between human operators and autonomous systems will likely include a diverse space of tools that include visual, aural, and tactile components. In all cases, there is the challenge of translating human goals into computer instructions (e.g., "solving" the long-standing "AI problem" of natural language processing), as well being able to

[564] See earlier discussion of the state-of-the-art in AI and these references: G. Bekey, *Autonomous Robots: From Biological Inspiration to Implementation and Control*, MIT Press, 2005; G. Weiss, editor, *Multiagent Systems*, Second Edition, MIT Press, 2013; S. Lucci and D. Kopec, *Artificial Intelligence in the 21st Century*, Mercury Learning and Information, 2013; E. Davis, "The singularity and the state of the art in artificial intelligence," presented at the Technological Singularity Ubiquity Symposium, *Ubiquity*, October 2014.

depict the machine's "decision space" in a form that is understandable by the human operator (e.g., allowing the operator to answer the question, *"Why did the system choose to take action X?"*)

6. *Predictability/control:* as autonomous systems increase in complexity, we can expect a commensurate decrease in our ability to both predict and control such systems; i.e., the "spectre of complacency in complexity." As discussed earlier, and to extent that, say, "deep learning" techniques will play a key role in shaping the "AI/software component" of autonomous systems, recall that there is a fundamental tradeoff between being able to achieve a given performance level (e.g., *AlphaGo*'s super-human ability to play Go) and to simultaneously know how that performance is achieved (e.g., the innate lack of access of the "outside world"—including the programmers—to *AlphaGo*'s internal "rule set" that determines *why* given moves are selected).

7. *Commercial-military focus disconnect:* while we have argued that many key innovations in AI, robotics, and autonomy are coming (and are likely to continue to come) from the commercial sector, the innovation thrusts in the commercial sector are not completely aligned with military needs: the problem-solving environments are typically "simpler" and easier to pre-specify; relegating "hard parts of problems" to humans to solve is more often an acceptable option; and, except for cyber-attacks, commercial systems rarely deal with intelligent adversaries.

Interoperability

DoD policy[565] mandates that IT and National Security Systems (NSS) shall be interoperable with existing and planned systems and equipment of joint, combined, and coalition forces, other U.S. Government departments and agencies, and nongovernmental organizations. Specifically, the policy requires DoD components to develop, acquire, test, deploy, and maintain ITs that are "interoperable and supportable with existing, developing, and proposed (pre-Milestone A) ITs through architecture, standards, defined interfaces, modular design, and reuse of existing IT solutions," and "are interoperable with host nation, multinational coalition, and federal, state, local, and tribal agency partners."[566] And DoD Directive 5000.01[567] requires that "systems, units, and forces shall be able to provide and accept data, information, materiel, and services to and from other systems, units, and forces and shall effectively interoperate with other U.S. Forces and coalition partners."

[565] CJCSI 6212.01F, Net Ready Key Performance Parameter (NR KPP), 21 March 2012.

[566] Ibid., p. 3.

[567] DoDD 5000.01, *The Defense Acquisition System*, 12 May 2003.

Since DoD unmanned systems have up until now been driven by quick-development timelines to accommodate operational requirements (and developed primarily for Service-specific needs), deployed systems have typically demonstrated only limited interoperability with other manned and unmanned platforms across Services.[568] While there are efforts to develop interoperability standards of message formats, architectures, and data protocols for unmanned systems (e.g., the NATO Standard Agreements, STANAG 4586, STANAG 4609, or the Joint Architecture for Unmanned Systems, JAUS),[569] they have thus far been used mainly to drive modular and re-usable designs of unmanned platforms and components, not facilitate operational machine-machine collaboration or cross-domain autonomy.[570] It is reasonable to expect that, as autonomous capabilities increase, and as the operational push towards integrating, say, air and ground vehicles also strengthens, there will be a concomitant impetus to develop a robust set of interoperability protocols. However, this will be impossible without a set of accepted set of standard T&E procedures for assessing autonomous unmanned system performance, something that DoD's acquisition currently lacks (see discussion in next section).

Trust

Research by the general HRI community has shown that *trust* plays a critical role in shaping an operator's interaction with an autonomous system.[571] However, trust is not an innate trait of the system; rather, it is a relative measure of how a human operator (or operators)—whose own performance depends, in part, on collaborating in some way with the system—experiences and perceives the behavior of the system; or, better, *how a human operator perceives the behavioral pattern of a system*. Trust is "an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability."[572] Though there are many alternative formulations of this basic description, attempts at a more formal definition, and lists of basic attributes needed, to varying degrees, for the creation of trust in a system, the underlying truth—and, as will be argued, a key reason why autonomy is so hard to certify—is that *trust is an abstraction that cannot be easily mathematized*. No

[568] *Unmanned Systems Integrated Roadmap: FY2013-2038*, U.S. Department of Defense: http://archive.defense.gov/pubs/DOD-USRM-2013.pdf.

[569] P. Durst and W. Gray, *Levels of Autonomy and Autonomous System Performance Assessment for Intelligent Unmanned Systems*, US Army Corps of Engineers, Engineer Research and Development Center, Geotechnical and Structures Laboratory, ERDC/GSL SR-14-1, April 2014.

[570] Ibid., p. 8.

[571] P. Hancock, et al., "Human-Automation Interaction Research: Past, Present, and Future," *Ergonomics in Design: The Quarterly of Human Factors Applications* 21, April 2013.

[572] J. Lee and K. See, "Trust in automation: designing for appropriate reliance," *Human Factors* 46, no. 1, 2004.

absolute measure of trust exists; rather, it is a relative measure that can be described only in terms of how an a priori level of trust has changed.

HRI research has also found that trust must be properly calibrated to ensure safe operation of an autonomous system. Too much trust can lead to abuse of the autonomous system; too little trust can lead to disuse of the autonomous system.[573]

Both the 2012[574] and 2016[575] Defense Science Board (DSB/2012 and DSB/2016, respectively) studies on autonomy have recognized the importance of trust in the development and deployment of autonomous systems. DSB/2012, in the context of summarizing the challenges inherent in the adoption of autonomy, states that for "commanders and operators in particular, these challenges can collectively be characterized as a *lack of trust* that the autonomous functions of a given system will operate as intended in all situations."[576] In DSB/2016, trust is a major issue around which much of the narrative is woven:

> Trust is complex and multidimensional. The individual making the decision to deploy a system on a given mission must trust the system; the same is true for all stakeholders that affect many other decision processes. Establishing trustworthiness of the system at design time and providing adequate indicator capabilities so that inevitable context-based variations in operational trustworthiness can be assessed and dealt with at run-time is essential, not only for the operator and the Commander, but also for designers, testers, policy and lawmakers, and the American public.[577]

Trust has been studied extensively both for human teaming[578] and in the general industrial robotics and automation literature,[579] and entails at least as many philosophies and approaches as does the literature on autonomy itself. For example,

[573] M. Desai, et al., "Effects of Changing Reliability on Trust of Robot Systems," in *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, Boston, Massachusetts, March 2012.

[574] *The Role of Autonomy in DoD Systems*, DoD DSB, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

[575] *Summer Study on Autonomy*, DoD, DSB, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016.

[576] Executive Summary, *The Role of Autonomy in DoD Systems*, 2012.

[577] Section 2, *Summer Study on Autonomy*, 2016.

[578] N. Stanton, editor, *Trust in Military Teams*, Ashgate Publishing Company, 2011.

[579] M. Dzindolet, et al., "The role of trust in automation reliance," *International Journal of Human-Computer Studies* 58., 2003.

Lee and See[580] summarize no less than 14 separate studies on how to define "trust" in automated systems. The key to all of these studies is a set of "dimensions" that describe the basis of trust. Examples include:[581]

- *Benevolence*: does the system support the mission and operator?
- *Directability*: can a system's actions be redirected by the operator?
- *False-alarm rate*: are a system's error rates known and acceptable?
- *Perceived competence*: does the operator believe the system can perform its assigned tasks?
- *Reliability*: the system has only a small probability of failing during a mission
- *Robustness*: how gracefully does the system respond to perturbations?
- *Understandability*: are the reasons behind a system's behaviors clear?
- *Utility*: do a system's actions add demonstrable value to a mission?
- *Validity*: is the system solving the correct set of problems?

Of course, just as for autonomy, for which there exist myriad taxonomies and dimensions (see figures 27, 28, and 29), there is no objective absolute measure of "trust." It is, at best, a relative concept, such that terms such as those that appear in the above list can be used to measure the *relative differences* between (e.g., "Is system A more or less 'reliable' than system B for a given set of tasks?"). If, and when, trust is established (by human operator, P, in a given system, S), it will *not* be the result of some "trust threshold" being exceeded (and measured in a mathematically precise manner). Rather, the trust will emerge over time, as P trains and works with S, and eventually determines—partly as a result of objective measures, and partly as a result of a subjective assessment of S's patterns of behavior—that S can adequately perform the set of tasks assigned to it. Even so, trust is more of a dynamic attribute of an ongoing series of human-machine collaborations than it is a static measure (that, once achieved, remains fixed). Relationships between human operators and robots can evolve and mature just as they do between humans.[582] They can strengthen, weaken, and transform (in unanticipated ways) over

[580] J. Lee and K. See, "Trust in automation: designing for appropriate reliance," *Human Factors* 46, no. 1, 2004.

[581] G. Palmer, A. Selwyn, and D. Zwillinger, "The Trust 'V': Building and Measuring Trust in Autonomous Systems," Chapter 4 in *Robust Intelligence and Trust in Autonomous Systems*, edited by R. Mittu, et al., Springer-Verlag, 2016.

[582] We do not have the space here to examine the detailed similarities and differences between trust of humans and trust of machines. While some similarities may be self-evident (as evidenced by the intuitive appeal of the list of basic dimensions of trust), there are

periods of time, and the degree of trust that an operator bestows on a machine, in general, depends on specific contexts and mission goals.

DSB/2016 identifies six key issues and barriers to establishing trust in autonomous systems:[583]

- *Sensing and thinking disconnect*: autonomous systems are likely to perceive the environment (using a variety of sensors and data sources not available to human teammates) and "reason" about it very differently from the way humans do (e.g., scene analysis derived from deep learning neural networks).
- *Lack of situational awareness*: even if an autonomous system is able to perform its tasks adequately for a fixed environment and when it is in its nominal state (and thereby, at least, provisionally, be deemed "trustworthy"), one of the key technical challenges for autonomy in general—namely, the ability of a system to self-inspect and to be generally aware of the environment, and any changes to it—is also an issue for establishing trust.
- *Predictability and directability*: autonomous systems must not only be able to inform their human teammates of any relevant information about their mutual environment (and to do so in an understandable way), but must also be able to anticipate events as they might unfold (e.g., it is hard to "trust" a system that may lead teammates into a dangerous situation because of an inability to predict it). In the event that something goes wrong (because of an action taken by the machine), the system must be able to both inform its teammates of the reasoning behind its action(s), and be amenable to redirection.
- *Commensurability of human-machine goals*: trust is difficult to achieve in human-machine teaming unless there is a mutual understanding of the goals of a mission.[584] DSB/2016 cites the example of how many of the commercial aircraft accidents that occurred in the 1990s resulted from a basic disconnect between human goals (e.g., stay on the glide slope during landing) and machine goals, in this case the flight computer (e.g., execute a go-around). As the software driving autonomous systems becomes more sophisticated, and

differences. For example, humans are generally more forgiving of trust breeches by other humans than they are of breeches by machines. Also, while humans may quickly "forgive" a breech if it is followed by a "quick confession" and/or explanation, by another human, the same is not true for a machine. Ref: R. Hoffman, et al., "Trust in Automation," *IEEE Intel. Sys.* 28, no. 1, 2013.

[583] Section 2, *Summer Study on Autonomy*, 2016.

[584] This is also an issue that lies at the heart of ethics issues surrounding the use of lethal autonomous weapons (discussed in a later section).

recedes more and more from "simple" outside inspection (not just from the developers, who may be using, say, neural-net-based algorithms, but from operators as well, who may erroneously "trust" that a system is driven by a published set of "high level" goals, but whose behavior, in reality, derives from a set of "low-level" goals that are invisible to the operator), the potential for disconnects between what an operator expects a machine to do and what the machine does obvious increases.

- *Human-machine interfaces*: to the extent that trust is a product of effective collaboration between humans and machines, traditional interfaces (such as mouse point-and-click) that slow rather than enhance real-time coordination and cooperation create barriers to developing trust. Ideally, human-machine teaming will rely on voice-command and dialogs powered by natural language processing algorithms, but such capabilities are beyond current state-of-the-art.
- *Adaptive machine learning*: since (full) autonomy requires that a machine be able to learn and adapt to changing environmental conditions, it is not only a difficult technical challenge on its own, but impacts both trust (i.e., how does the trust that has been "earned" by a system when operating in a fixed environment translate to different environments, and/or different machines as new team members?)[585] and verification and validation, since the conditions under which verification and validation is initially performed may no longer be valid.

Establishing and maintaining trust in an autonomous system requires a continual feedback between human (developers, operators, and commanders) and machine during the entire lifecycle of the system; and includes not just development and experimentation, but operational contexts as well. Palmer, et. al.,[586] introduce a basic framework that combines a list of trust attributes with factors that describe autonomy, and show how it can be used to modify the systems engineering model—built into the existing DoD acquisition process—to enhance the ability to "build in" trust. We will touch on aspects of this framework when we discuss the inherent limitations of the test and evaluation (T&E) and verification and validation of autonomous systems in the following section.

[585] Palmer, et al. ("The Trust 'V': Building and Measuring Trust in Autonomous Systems," Chapter 4 in *Robust Intelligence and Trust in Autonomous Systems*, edited by R. Mittu, et al., Springer-Verlag, 2016), add systems-of-systems integration to their list of "dimensions" that describe the basis of trust, by which they mean the undesirable emergent behaviors that may emerge when multiple systems (autonomous or not) are combined.

[586] Ibid., Section 4.5.

Acquisition process

The DoD Acquisition Process (DAP) is one of three procurement processes that make up the Defense Acquisition Management System (DAMS), and is implemented by DoD Instruction (DoDI) 5000.02.[587] Acquisition programs consist of a series of milestone reviews and other decision points that authorize entry into a new program phase. This instruction defines the policies that govern the DAS and forms the management foundation for all DoD programs that include weapon systems, services, and Automated Information Systems (AIS). DoD Instruction 5000.02 also establishes a program management framework for translating user needs and technology opportunities into stable, affordable and well-managed acquisition programs; and identifies specific statutory and regulatory reports and other information requirements for each milestone and decision point (defined below). Since there are many more layers of detail to the DAS than space permits,[588] we focus our discussion only on those aspects that directly impact the subject matter of this report.

Figure 35 shows a generic schematic of the DAP. Two notable time scales are: (1) 91 months, on average, from start of an Analysis of Alternative (AoA) study to Initial operational capability (IOC), and (2) historically, Information Technology (IT) programs have averaged 81 Months. Intermixed on this main timeline are two other telling time-scales: an average of 3 years to accommodate personnel rotation, and a roughly two-year cycle of technology change.

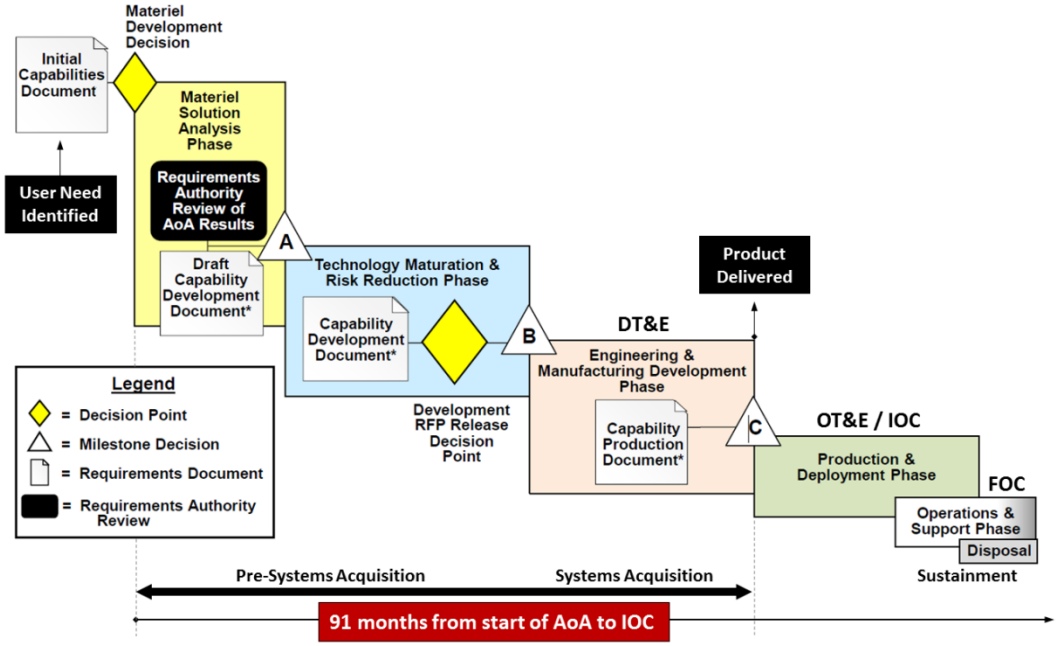
The DAP is predicated on a *user need* being identified (i.e., the first requirements authority review block shown center-left in figure 35). DoDI 5000.02 defines this first step in the acquisition process as the "capability needs and acquisition management systems shall use Joint Concepts, integrated architectures, and an analysis of doctrine, organization, training, materiel, leadership and education, personnel, and facilities (DOTMLPF) in an integrated, collaborative process to define needed capabilities to guide the development of affordable systems."[589]

[587] DoD Instruction 5000.02, *Operation of the Defense Acquisition System*, January 7, 2015: http://acqnotes.com/wp-content/uploads/2014/09/DoD-Instruction-5000.02-Operations-of-the-Defense-Acquisition-System-7-Jan-2015.pdf.

[588] The most recent release of the *Defense Acquisition Guidebook* (16 Sep 2013) contains 1248 pages: acc.dau.mil/docs/dag_pdf/dag_complete.pdf.

[589] Enclosure 2, DoDI 5000.

Figure 35. Generic DoD acquisition process timeline



Ref: DoD, Defense Science Board, *Policies and Procedures for the Acquisition of Information Technology*, March 2009.

The generic DAP (as depicted in figure 35) consists of five core phases:[590]

1. *Materiel Solution Analysis* (MSA): assesses potential solutions for a needed capability as defined in an Initial Capabilities Document (ICD), and requires an Analysis of Alternatives (AoA) study (the purpose of which is to evaluate the mission effectiveness, operational suitability, and estimated life-cycle cost of alternative solutions for meeting an ICD-specified mission capability).
2. *Technology Maturation & Risk Reduction* (TMRR): the goal of which is to produce a working prototype that allows for a basic assessment to be made of the technology, risk, and design. This phase includes competitive prototyping of system elements, refinement of requirements, and the development of the functional and allocated baselines of the end-item system configuration ("allocated baselines" refers to how system function and performance requirements are allocated across lower level configuration items). The objective is to develop a sufficient understanding of a solution to

[590] *Defense Acquisition Portal*: https://dap.dau.mil/aphome/das/Pages/Default.aspx

allow sound business decisions on starting a formal acquisition program in the Engineering & Manufacturing Development (EMD) phase.

3. *Engineering & Manufacturing Development* (EMD)*:* in which the prototype enters a bona-fide developmental stage, and interim designs can be tested (via developmental and operational tests, and live-fire test and evaluation), and the final prototype undergoes critical design review. The EMD phase starts after a successful milestone B (see figure 35) and pre-EMD review and is considered the formal start of the actual system program. The goal is to complete the development of a system (or increment the capability of an existing system), complete full system fabrication and integration, and test and evaluate the system before moving into the next phase.
4. *Production and Deployment* (PD): during which a system that satisfies an operational capability is produced and deployed to the end user. The phase begins after a successful milestone C review and the EMD Phase is complete, and includes both Low-Rate Initial Production (LRIP), in which a small-quantity set of systems is produced for Initial Operational Test and Evaluation (IOT&E), and Full-Rate Production and Deployment (FRP&D).
5. *Operations and Support* (O&S): during which a system is used and supported by users in the field. The main focus is the general cost-effective support to sustaining a system (e.g., maintaining capabilities, logistics, and upgrades). The last component of O&S (and in the overall life-cycle of a system) is the system's disposal after it has reached the end of its useful life.

Challenges: (general) technology related

Acquisition challenges, particularly for IT systems and general weapons systems that include a heavy coupling between hardware and software, have been known—and debated—for decades.[591] However, despite numerous attempts by various stakeholders to address these challenges, the generic acquisition process (at least on the traditional institutional level) remains effectively unchanged. Whatever progress has been made in recent years derives more from *workarounds* instituted by DoD to facilitate "rapid acquisition" of systems,[592] rather than wholesale changes applied to stovepiped processes of the DAP itself.

[591] J. Merritt and P. Sprey, "Negative marginal returns in weapons acquisition," *in American Defense Policy*, Third Edition, edited by R. Head and E. Roppe, John Hopkins Univ. Press, 1973.

[592] Examples include: U.S. Air Force Rapid Capabilities Office, the U.S. Army's Asymmetric Warfare Group and Rapid Capabilities Office, DoD's Strategic Capabilities Office, and, most recently, SecDef's Ashton Carter's Defense Innovation Unit Experimental (DIUx). Ref: B.

Nonetheless, some recent progress has been made. For example, the 2009/2011 National Defense Authorization Acts (NDAA/Sec 804), mandated a new IT acquisition process.[593] This led to multiple Defense Science Board (DSB) Task Force (TF) studies of general issues of the acquisition process, which collectively concluded that:[594]

- Oversight process not aligned with rapid acquisitions (favors large, high-level oversight)
- Systems take too long to deliver and inconsistent with technology cycles
- Overly detailed requirements inconsistent pace of technology change and need for rapid delivery
- Inadequate metrics to assess IT-based systems performance
- Testing integrated too late and serially
- Cyber-security is inadequately managed during the acquisition process
- Significant cultural impediments to change

A notable *absence* in any of the above cited DSB/TF studies is any explicit mention of autonomy.

DoDI 5000.02 canceled the interim 5000.02 version (issued Nov 2013, and which itself replaced DoDI 5000.01/2007), and explicitly implements policies and practices in the Better Buying Power (BBP) initiative.[595] The BBP is designed to achieve greater efficiency through affordability, cost control, elimination of unproductive processes and bureaucracy, and promotion of competition; and to incentivize productivity and innovation in industry and government. Specifically, DoDI 5000.02 generalizes the basic timeline as depicted in figure 35 by replacing the single "generic" model with multiple system-*type*-specific program structure models (note that the five core phases outlined earlier—MSA, TMRR, EMD, PD, and O&S—remain unchanged,

Fitzgerald, A. Sander, J. Parziale, *Future Foundry: A New Strategic Approach to Military-Technical Advantage*, Center for a New American Security, 2016.

[593] R. Pontius, Dir, C2 OUSD (Acquisition, Technology & Logistics), "Acquisition of Information Technology," March, 2012.

[594] DSB TF on Improvements to Services Contracting, March 2011; DSB TF on DoD Policies and Procedures for the Acquisition of IT, March 2009; DSB TF on Creating a DoD Strategic Acquisition Platform, April 2009; and DSB TF on Fulfillment of Urgent Oper. Needs, July 2009.

[595] Under Secretary of Defense for Acquisition, Technology and Logistics, or USD(AT&L), Memo, Implementation Directive for Better Buying Power 3.0: Achieving Dominant Capabilities Through Technical excellence and Innovation, April 2015: http://www.acq.osd.mil/fo/docs/betterBuyingPower3.0(9Apr15).pdf.

although the numbers and timing of certain milestone decisions and decision points depend on the specific model):[596]

- Model 1: *Hardware intensive*, which is the classic timeline that has existed in some form in all previous versions of this instruction; e.g., major weapon platforms.
- Model 2: *Software intensive,* that describes a model of a program that is dominated by the need for a complex—typically defense unique—software system that will not be fully deployed until several "builds" have been completed.
- Model 3: *Incrementally deployed software intensive*, which is intended mainly for Defense Business Systems (DBS)—i.e., information systems other than a national security system (NSS) operated by, for, or on behalf of DoD; financial systems, mixed systems, financial feeder systems). The model also applies to upgrades of some command and control systems, and weapons systems software (for which deployment of full capability occurs in multiple increments as new capabilities are developed and delivered, nominally in 1- to 2-year cycles).

  Model 3 applies specifically to cases where commercial off- the-shelf software are acquired and adapted for DoD applications. However, DoDI 5000.02 cautions against misuse of the model, in that, as currently formulated, there is the potential for the process to be overwhelmed with frequent milestone or deployment decision points and associated approval reviews.[597]
- Model 4: *Accelerated acquisition*, which is designed for cases where schedule considerations dominate over cost and technical risks. The model accommodates compression (and/or total elimination) of distinct phases of the DAP, and is intended to be used when a higher-risk acquisition program is required to respond to some "technological surprise" by a potential adversary. (There is a variant of this model that includes procedures for dealing with urgent needs that can be fulfilled in less than 2 years.[598])

[596] DoD Instruction 5000.02, *Operation of the Defense Acquisition System*, January 7, 2015: http://acqnotes.com/wp-content/uploads/2014/09/DoD-Instruction-5000.02-Operations-of-the-Defense-Acquisition-System-7-Jan-2015.pdf.

[597] Model 3 is further distinguished from Model 2 by the inclusion of multiple acquisition increments that facilitate rapid delivery of capability. Each increment is assumed to provide a part of the overall required program capability, and may have several limited deployments. Each deployment results from a specific build and providing the user with a mature and tested sub-element of the overall incremental capability. (It is assumed that several builds and deployments may be necessary to satisfy requirements for an increment of capability.)

[598] Enclosure 13 to DoDI 5000.02, pp. 143-152.

- Model 5: *Hybrid-A – hardware dominant concurrent with software*, which is a model intended to be used for the acquisition of major weapons systems that combine the simultaneous development of hardware and software. While the overall schedule is defined largely by the design, fabrication, and testing of physical prototypes, software development dictates the pace of program execution, and must be tightly integrated and coordinated with hardware development decision points.

  The Hybrid-A model assumes that software development is organized into a series of testable software builds (which, in turn, also assumes that software functional capability development maturity criteria and technical performance criteria both exist and are well-defined).

- Model 6*: Hybrid-B – software* dominant concurrent with hardware, which is the software-intensive complement to Model 5, and incorporates the same incremental software fielding policy as defined for Model 3. DoDI 5000.02 suggests that while the Hybrid-B model may be complex to plan and execute successfully, of the six variants, it may be the most logical way to structure the acquisition program. Figure 36 shows a schematic of the Hybrid-B model.

IT Box

In addition to these acquisition models—and to more explicitly take advantage of emerging commercial information technology—in 2014 the Chairman of the Joint Chiefs modified the Department's Joint Capability Integration and Development System (JCIDS) by introducing the *IT Box*.[599] The "IT Box" is designed to delegate authorities to specifically support the more rapid timelines necessary for IT capabilities through the DAP, and is named after the four sides of an organizational template that need to be defined: (1) the organization that will provide oversight and management of the product; (2) the capabilities required; (3) the cost for application and system development; and (3) the costs for system enhancements and integration. The *IT Box* can lead faster timelines for IT programs because the sponsor's organization is not required to return to the Joint Requirements Oversight Council (JROC) for approval of any changes to requirements unless the *IT Box* parameters are exceeded by prescribed thresholds. However, in terms of its ability to streamline the general acquisition of autonomy-enabling software and other innovative technologies, the *IT Box* suffers from the same fundamental limitation that applies to other acquisition models; namely, it does not accommodate the unique technical challenges of the design, development, testing, and accreditation of autonomous systems (as discussed in the next section).

[599] *Senate Armed Services Subcommittee on Readiness and Management Support Hearing*, February 27, 2014: https://www.insurancenewsnet.com/oarticle/Senate-Armed-Services-Subcommittee-on-Readiness-and-Management-Support-Hearing-a-466703#.UxXKQ_RdWV4.

Figure 36. Schematic of acquisition Model 6 (hybrid-B: software dominant concurrent with hardware)

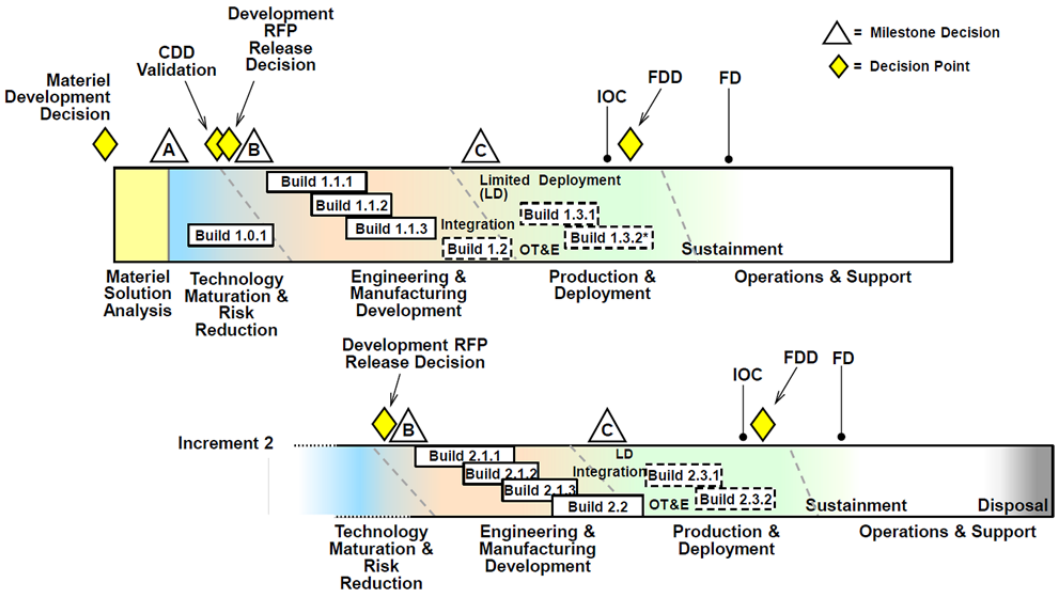


Figure 8 in Department of Defense, Instruction 5000.02, January 7, 2015; CDD=capability development document; FD = full deployment; FDD = full deployment decision; IOC = initial operational capability.

While DoDI 5000.02 emphasizes the special risks (to cost and schedule) that necessarily accompany all highly integrated complex software and hardware development processes, it does not provide any guidance on how to mitigate those risks (apart from asserting that risks "must be managed throughout the program's life cycle and will be a topic of special interest at all decision points and milestones" [600]).

Taking a more "bird's eye" view of the technological challenges facing DoD (beyond just the problems inherent in the acquisition process itself, but still directly addressing them), is a recent report[601] issued by The Center for a New American

[600] DoDI 5000.02, p. 15.

[601] B. Fitzgerald, A. Sander, nd J. Parziale, *Future Foundry: A New Strategic Approach to Military-Technical Advantage*, CNAS, Dec 2016: https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Report-FutureFoundry-final.pdf.

Security's (CNAS's) *Future Foundry* project (CNAS/FFP).[602] The report recommends that DoD take a new strategic approach—an "optionality strategy"—that emphasizes developing and sustaining technological advantage by expanding the available range of military and technical options across a more diverse portfolio of capabilities and concepts: DoD's "core acquisition challenge is not that its current system is fundamentally flawed, but that the department has many technology needs that the system is not designed to meet."[603] The main narrative of CNAS/FFP's report is woven around its delineation of four core capability segments: (1) *military unique systems with constrained competition*, which includes the traditional low-production/high-capital-investment weapons systems such as aircraft carriers and submarines, and whose suppliers are defense specialists; (2) *military unique systems with viable competition*, which includes most systems other than the large-scale examples included in the first segment (e.g., combat aircraft, armored vehicles, and unmanned systems); (3) *military adapted commercial technology*, the few current examples of which have all come from sources outside DoD's traditional acquisition pipeline[604] (and which refers generally to any technology that may be rapidly developed and deployed by leveraging emerging commercial technologies); and (4) *purely commercial technology*, which includes all commercial off-the-shelf purchases (and which, though accommodated for by the existing acquisition process,[605] is both underutilized and still reliant upon generating a stovepiped "military unique requirements" document prior to purchase).

Notably, the only capability segment not covered by any existing acquisition processes is the third, which CNAS/FFP's report identifies as the most promising opportunity for "new entrants." Noting that the existing requirements process has been optimized for developing large-scale, long-term, military-unique weapon systems—and is ill-equipped for dealing with fast-paced innovation—the report recommends that the DoD add multiple new acquisition pathways toward finding and introducing a wider, more diverse, range of technologies.

[602] The CNAS/FPP focuses on developing strategies that foster collaboration between DoD and potential partners from multiple commercial industry sectors. Ref: B. Fitzgerald, A. Sander, and J. Parziale, *Future Foundry: A New Strategic Approach to Military-Technical Advantage*, CNAS, Dec 2016

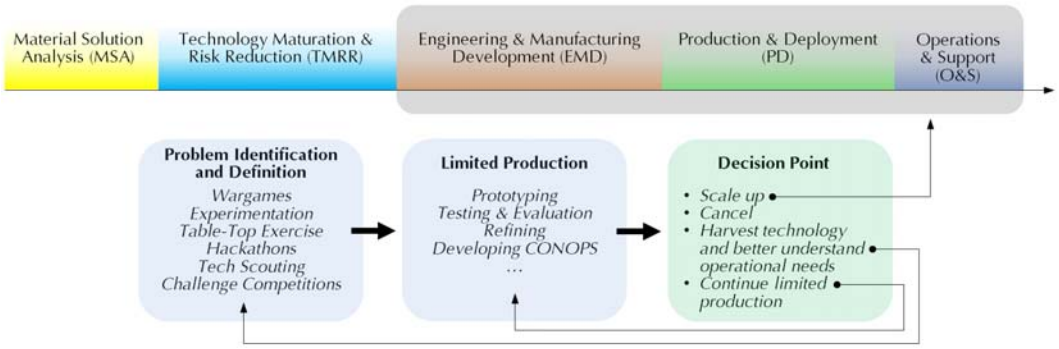
[603] Ibid., p. 22.

[604] The *Defense Innovation Unit Experimental* (DIUx) has been established, in part, to facilitate the discovery and development of precisely these kinds of outside-the-normal-acquisition-pipeline capabilities and technologies. Ref: https://www.diux.mil/.

[605] DoD has authority to acquire commercial technology under the policy defined in Part 12 (Acquisition of Commercial Items) of the Federal Acquisition Regulation (FAR): https://www.acquisition.gov/sites/default/files/current/far/pdf/FAR.pdf.

For example, one new pathway—forged as a mirrored complement to existing practices in the EMD, PD, O&S phases, and thereby effectively bypassing the traditional requirements phase (see figure 37)—is simply to incorporate some of the same methods for discovering and identifying promising technologies that are now in place only as workarounds to the acquisition system (e.g., as part of the aforementioned DIUx). Such methods include wargaming, table-top exercises, and challenge competitions. Systems spawned via this new pathway would necessarily be of limited production, with multiple simultaneous prototype variants serving as examples of possible capabilities about which DoD can then make better informed decisions (including the option of harvesting some of the developed concepts and/or technologies for use in existing systems). Moreover, integrating developmental and operational T&E into the early stages of the new pathway (particularly for systems that transition into full production) will militate some of the unique technical challenges facing the development of autonomous systems.

Figure 37. An addition acquisition pathway to accelerate adoption of innovative technology



After figure on page 26 in B. Fitzgerald, A. Sander, and J. Parziale, *Future Foundry: A New Strategic Approach to Military-Technical Advantage*, CNAS, Dec 2016.

Challenges: autonomy related

All five phases of the DAP (whether generic, as shown in figure 35[], or defined by the six system-*type*-specific models defined in DoDI 5000.02) contain elements that touch upon—indeed, as will be argued below, touch *deeply* upon—autonomous weapons systems (AWSs). For example (see figure 35):[606] (1) the MSA phase requires an assessment of potential solutions for a stated need, and a specification of

[606] DoD Instruction 5000.02, *Operation of the Defense Acquisition System*, January 7, 2015

program goals for any needed development of critical enabling technologies; (2) the TMRR phase requires Test and Evaluation (T&E) and risk assessment plans; (3) the EMD phase involves a tight integration of hardware, software, and human systems, and requires a demonstration of full system integration, interoperability, supportability, safety, and utility (and involves no less than four separate technical reviews: a critical design review, test and readiness review, a system verification Review, and a technology readiness assessment); and (4) the PD and O&S phases require an updated T&E plan and risk assessment, and modifications and upgrades to fielded systems and data collection for an in-service review, respectively.

However, none of these requirements explicitly reference the unique characteristics of AWS. Neither the requirements nor sequencing (nor even the detailed steps) of individual phases take into account the fact that the acquisition of *autonomous* weapons entails testing and evaluation procedures distinctly different from conventional systems, even those with an intensive software focus (i.e., a major distinctive element may be appreciated, intuitively, by reflecting on the difference between "conventional" operating instructions for automated hardware systems and AI-derived behavioral-logic necessary to govern autonomous vehicles).

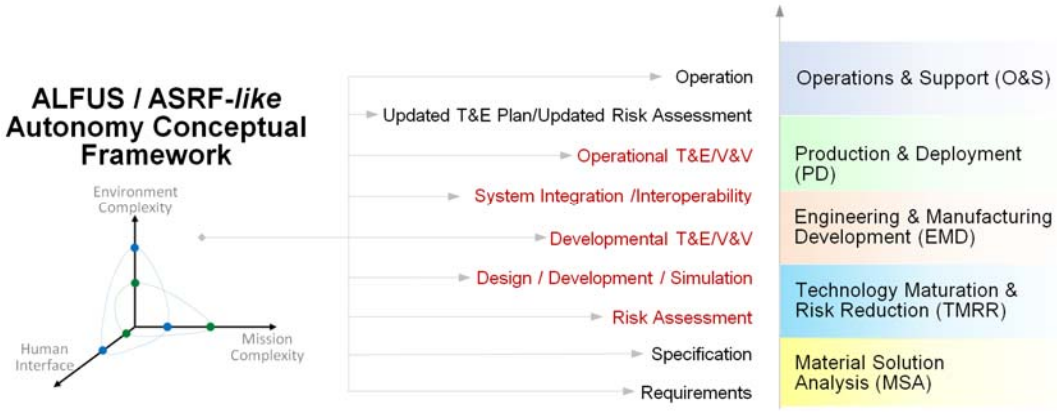
Figure 38 illustrates how an ALFUS-like conceptual framework of autonomy can be used to help support various components of the acquisition process. The core elements (highlighted in red) involve risk assessment, modeling and simulation, and developmental and operational T&E and V&V, a discussion of which we turn to next.

DoD's Directive (DoDD) 3000.09 ("Autonomy in Weapon Systems") requires that weapons systems:[607]

- Go through "rigorous hardware and software verification and validation (V&V) and realistic system developmental and operational test and evaluation (T&E), including analysis of *unanticipated emergent behavior* resulting from the effects of complex operational environments on autonomous or semiautonomous systems"
- "Function as anticipated in realistic operational environments against *adaptive adversaries*"
- "Are sufficiently robust to minimize failures that could lead to *unintended engagements*"

[607] Enclosures 2 and 3 of DoD Directive 3000.09 (*Autonomy in Weapon Systems*, Nov 2012) address T&E and V&V issues, and general review guidelines, respectively.

Figure 38. Schematic of how an ALFUS-like autonomy conceptual framework cam help support the acquisition process



Expanding on V&V and T&E, DoDD 3000.09 requires that they must "assess system performance, capability, reliability, effectiveness, and suitability under realistic conditions, including *possible adversary actions*, consistent with the *potential consequences of an unintended engagement* or loss of control of the system."

Furthermore, in regard to the human-machine interface (for autonomous and semi-autonomous weapon systems), it must be:

- "Readily *understandable* to trained operators"
- "Provide *traceable feedback* on system status"
- "Provide clear procedures for trained operators to *activate and deactivate system functions*"

The italics (added by the author of this memorandum) are meant to highlight that the "Devil is in the details!" It is one thing to set policy; it is quite another matter to apply existing (and/or develop new) concepts and methods to ensure that policy requirements are actually met. Given the lack of transparency of "narrow AI" methods (e.g., recall the current inability of otherwise at-least-human-level-performing neural-net-based algorithms to "explain" their methods even to the designers themselves), in general, and the inevitability of self-organized emergence in complex adaptive systems (recall that all sufficiently complex systems are guaranteed to display fundamentally unpredictable behaviors), we can anticipate that few if any existing DoD practices (e.g., DAP, in general, and T&E and V&V, in particular; see discussion below) are adequate in their current form to accommodate, much less ensure the viability of, the policies set forth in DoDD 3000.09. We will revisit this key issue in a later section.

T&E / V&V and Accreditation (VV&A) Challenges

According to DoD Instruction 5000.61—for DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)—accreditation, verification, and validation are defined as follows:[608]

- *Accreditation:* the official certification that a model or simulation and its associated data are acceptable for use for a specific purpose; "certification" is defined as the formal acknowledgement that a system (or program) meets a specific set of requirements and has passed the T&E/VV&A process.[609]

- *Verification:* the process of determining that a model or simulation implementation and its associated data accurately represent the developer's conceptual description and specifications; i.e., verification effectively answers the question, *"Did we build the system correctly?"*

- *Validation:* the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model; i.e., validation answers the question, *"Is the system the right solution to the problem?"*

Current engineering methods for T&E/VV&A of hybrid hardware-software systems lack sufficient fidelity and robustness to deal with highly complex, software intensive systems. The most difficult and challenging component of autonomous weapon systems to certify is the AI/machine-learning/adaptive *software* that is embedded within them.[610] The DSB/2012 report on autonomy states:[611]

> Unlike many other defense systems, the critical capabilities provided by autonomy are embedded in the system software. However, the traditional acquisition milestones for unmanned systems, often along with the focus of the development contractor, are dominated by hardware considerations. Autonomy software is frequently treated as

[608] Department of Defense Instruction 5000.61, *DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)*, USD(AT&L), Dec 9, 2009: http://www.dtic.mil/whs/directives/corres/pdf/500061p.pdf.

[609] *Defense Acquisition Handbook*, Chapter 4: Systems Engineering, https://acc.dau.mil/docs/dag_pdf/dag_complete.pdf.

[610] *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under SecDef for Acquisition, Technology and Logistics, June 2016

[611] *The Role of Autonomy in DoD Systems*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

> an afterthought or assumed to be a component that can be added to the platform at a later date—independent of sensors, processing power, communications and other elements that may limit computational intelligence.

T&E of critical software for conventional (i.e., non-autonomous) systems has been estimated to cost *seven times* that of software development costs.[612] The associated T&E costs for software driving *autonomous* behavior will be at least this great. Also, current practice calls for *full* system tests, which will become increasingly infeasible as autonomous systems attain ever greater levels of self-governance. In designing and certifying conventional software, one typically needs only to address the issues involved in answering *what* (the software does) and *when* (the software does it). Since autonomous systems must make their own decisions, autonomy inevitably introduces the more complex *why* (the software chooses to do something).

In a 2011 memo, the SecDef designated autonomy and human-machine interface systems as two (of 7) priority S&T investments.[613] In response, the Assistant SecDef, Research and Engineering (R&E) set up the Autonomy Community of Interest (COI) and identified four challenge areas, with a designated working group (WG) for each:[614]

- *Human/ Autonomous Systems Interaction and Collaboration*
  - Robust cognitive and neurological or other models that can model human interaction and teaming with autonomous systems beyond fairly narrow applications.
  - Integration of autonomy, artificial intelligence and human cognitive, or other human models
  - Optimized trust in automation/transparency
  - Principled control station human factors engineering
  - Advanced feedback interfaces to maximize common perception between human(s) and agent(s)

[612] C. Hang, P. Manolios, and V. Papavasileiou, "Synthesizing cyber-physical architectural models with real-time constraints," *Computer Aided Verification*, Springer Berlin Heidelberg, 2011.

[613] Memo, *Science and Technology (S&T) Priorities for Fiscal Years 2013-17 Planning*, Secretary of Defense, 19 April 2011: http://www.acq.osd.mil/chieftechnologist/publications/docs/OSD%2002073-11.pdf.

[614] *Autonomy Research Pilot Initiative*, DOD Priority Steering Council (PSC), ASD (R&E). Nov 2012: http://auvac.org/uploads/publication_pdf/Autonomy%20Research%20Pilot%20Initiative.pdf.

  - Secure communication between human(s) and agent(s) and understanding of intent and actions of human team members, adversaries and bystanders
  - Advanced control system interfaces

- *Scalable Teaming of Multiple Autonomous Systems*
  - Shared problem solving/reasoning between agents
  - Shared perception between agents
  - System health management/attrition management
  - Secure communication between multiple agents
  - Scalable collaboration among heterogeneous teams

- *Machine Reasoning, Perception, and Intelligence*
  - Data-driven analytics
  - Sensor/data decision models
  - Advanced algorithms to enable robust operations in unstructured environments including machine learning
  - Contingency-based control strategies
  - Adaptive guidance and control integration with higher level reasoning, decision making, learning.
  - Domain management (e.g., air, sea, land) and mission control
  - Integrated contextual decision making
  - Cognitive, intelligent, and adaptive computing paradigm/platform

- *T&E and V&V (TEVV)*
  - Live and simulation test beds for
    - Human-agent teaming
    - Operation in complex, contested environments
    - Controlled, coordinated actions by multiple agents
  - Methods and strategy…
    - To test and evaluate autonomous systems/subsystems
    - To Test and Evaluate Human-Agent Interfaces
    - For Validation and Verification of Computer Models/Logic
    - For algorithms, and integrated software tools

In 2013, the Autonomy COI TEVV (ATEVV) working group (ATEVV/WG) held several workshops to identify the core challenges facing the T&E of autonomy. The results of this effort, published in a memo by Assistant SecDef(R&E),[615] included four general autonomy-specific challenges and six specific technical gaps that collectively identify the ATEVV/WG's recommended changes to the current V&V paradigm. Challenges include:

1. *State-space complexity:* autonomous systems, by their nature, possess a near infinite number of possible system "states" that must be tested. The algorithmic decision space is either non-deterministic (i.e., output cannot be predicted because of multiple possible outcomes for each input), or intractably complex; in either case, it is not possible to conduct an exhaustive search of all possibilities. Thus, the existing requirements-driven design and T&E/VV&A process—in which T&E is built around the requirements that define the desired system response(s) for all conditions—is ill-equipped for dealing with autonomous systems.

2. *Environmental complexity:* while the benefits of autonomy derive principally from the ability of autonomous systems to function in unknown and/or untested environments, testing for desired behaviors only exacerbates the "state-space complexity" problem. Since the behavior of an autonomous system cannot be specified (much less tested and certified) in situ, but in concert with interaction with a dynamic environment via sensors, effectors, and communications links, the explicit specification of all combinations of the system inputs/outputs and environmental variables is combinatorically impossible.[616]

3. *Emergent behavior:* the appearance of emergent behavior is a well-known property of all complex adaptive systems (CAS). To the extent that unscripted behaviors of autonomous systems—whether they are individual systems or are components of larger swarms—derive from CAS-like elements (as discussed earlier), we can expect novel or unexpected behavior to arise naturally and unpredictably in certain dynamic situations. Existing

[615] Memo, Assistant SecDef(R&E), *Autonomy Test and Evaluation, Verification and Validation Technology Investment Strategy: 2015-2018*, 12 June 2015: http://www.defenseinnovation marketplace.mil/resources/OSD_ATEVV_STRAT_DIST_A_SIGNED.pdf.

[616] G. Zacharias, *Advancing the Science and Acceptance of Autonomy for Future Defense Systems*, presented to the House Armed Services Committee, Subcommittee on Emerging Threats and Capabilities, U.S. House of Representatives, 19 Nov 2015: http://docs.house.gov/ meetings/AS/AS26/20151119/104186/HHRG-114-AS26-Wstate-ZachariasG-20151119.pdf.

T&E/VV&A practices do not have the requisite fidelity to deal with emergent behavior.[617]

4. *Human-machine dynamics:* the operational effectiveness of autonomous systems (regardless of the details of whatever method is used to measure it) ultimately depends on the dynamic interplay between the human operator and the machine(s) in a given environment, and how the system responds, in real-time, to changing operational objectives as the human adapts to dynamic contexts—an interplay that plays an even more important role in human-machine *teaming*. The efficacy of this human-machine interplay, from the human operator's point of view, is driven by the degree of "trust" that the operator has in the behavior and performance of the machine, a basic point stressed by the Defense Science Board's most recent study on autonomy.[618] However, since "trust" is not an innate trait of the system, and the current T&E/VV&A process is designed to test systems in closed, scripted environments, the human-machine dynamics is not naturally accommodated. On an even simpler level, the human component will require *in*– and/or *on*–the–loop experimentation, which again limits the effective dimensionality of the test space.

   "Trust" also entails grappling with the issue of *experience* and/or *learning*: at the high end of autonomous systems will be those that are able to accrue and learn from "operational experience," both in real-time, adapting to changing conditions as the mission unfolds, and over longer periods of time. Such systems cannot be certified monolithically, in one "check in the box" moment of time. Rather, they will require periodic retesting and recertification, the periodicity of which is (some as yet undetermined) function of the system's history and "experience."

The ATEVV/WG also identified six specific autonomy-related technical gaps in existing V&V practice:

[617] Certification is applied at the system level; i.e., sub-systems are certified only as parts of a system. Verification is applied only after system integration (by which time any errors that are identified may be too costly to fix), and validation of requirements occurs only at the end of development (by which point, a problem in requirements may invalidate the entire system). Ref: *Test and Evaluation Management Guide*, Sixth Edition, Department of Defense, December 2012; Department of Defense Instruction 5000.61, DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A), USD(AT&L), Dec 9, 2009: http://www.dtic.mil/whs/directives/corres/pdf/500061p.pdf.

[618] *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016.

1. *Lack of Verifiable Autonomous System Requirements:* there are no common, clear, and consistent requirements for systems that include autonomy requirements (particularly in regard to assumptions about the environment, Concept of Operations (CONOPS), interoperability, and communication; and no universally-agreed-upon Measures of Effectiveness (MoEs), Measures of Performance (MoPs), or other metrics.

2. *Lack of Modeling, Design, and Interface Standards:* there are no current standardized modeling frameworks for autonomous systems that span the whole system lifecycle (R&D through T&E). As a consequence, there is currently a significant gap in traceability that exists between capabilities (whether implemented in conventional non-autonomous systems or systems that include features such as adaptivity, nonlinearity, and/or learning) and the requirements they are designed to meet.

3. *Lack of Autonomy Test and Evaluation Capabilities:* there is a lack of T&E ranges, testbeds, and skillsets for handling dynamic learning and adaptive systems. The innate complexity of autonomous systems make it impossible to test these systems under all possible conditions (and accounting for diversity of environments and human-system interactions).

4. *Lack of Human Operator Reliance to Compensate for Brittleness:* while the burden of decision making under uncertainty currently lies solely under the purview of human operators, as systems evolve from exhibiting relatively predictable automated behaviors to those that are more complex and unpredictable, it will become increasingly difficult for human operators to understand and respond appropriately to the decisions made by autonomous systems. Existing V&V practices do not sufficiently factor in human-machine interfaces, human performance characteristics, and requirements for human operator training.

5. *Lack of Run Time V&V during Deployed Autonomy Operations:* The human operator currently acts as the ultimate arbiter and "fail safe" component to unmanned weapons systems. However, as autonomy increases (however it is defined), there will be a commensurately decreasing reliance on human intervention. The existing V&V process includes no mechanisms to ensure that autonomous systems that have not been fully tested (exhaustive testing for all possible conditions is something that is, in principle, impossible to achieve for any fully autonomous system) can be successfully deployed—and trusted to perform—in operational environments.

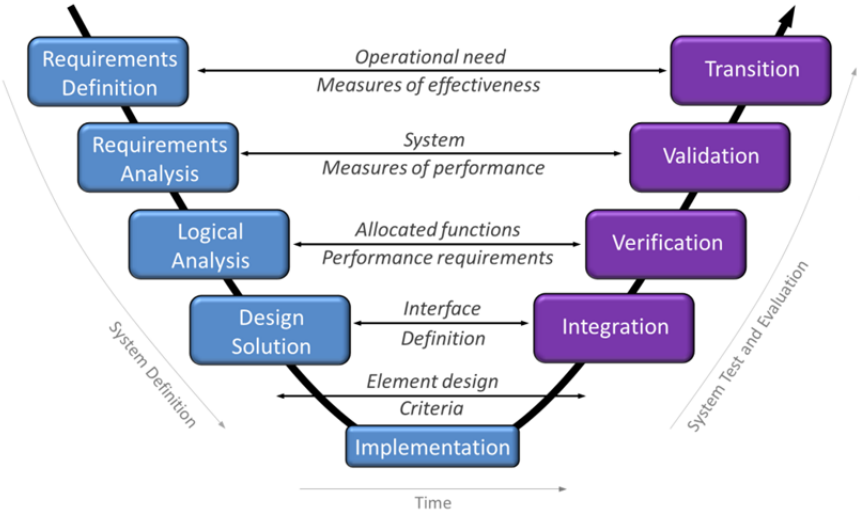
6. *Lack of Evidence Re-use for V&V:* The current practice of relying on a system's "past (tested) performance" (including archived "failures" in specific contexts) to adjudicate acceptable levels of safety, security, performance, and risk assumes that any lessons learned from the performance of a deployed system will apply also to any *similar systems*. This assumption no longer holds true for autonomous systems.

Large-scale complex engineering systems are typically managed using the "V" (or "Vee") model:[619]

> Various life cycle models such as the waterfall, spiral, Vee, and agile development models are useful in defining the start, stop, and activities appropriate to life cycle stages. The Vee model is used to visualize the system engineering focus, particularly during the concept and development stages. The Vee highlights the need to define verification plans during requirements development, the need for continuous validation with the stakeholders, and the importance of continuous risk and opportunity assessment.

The V-model is so-called because of how the stages of the model are usually visually depicted (see figure 39). In the figure (which can be compared to the diagram depicting the different stages of the acquisition process; see figure 35), time flows from left to right in a path that starts (on the left leg) with the requirements definition (at top left), moving to analyses and design solutions that provide detailed specifications for systems and assemblies.

Figure 39. Classic "V" systems engineering model



After figure 1 in Memo, Assistant SecDef(R&E), Autonomy Test and Evaluation, Verification and Validation Technology Investment Strategy, June 2015.

[619] C. Haskins, editor, *Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities*, International Council on Systems Engineering (INCOSE), June 2006: http://disi.unal.edu.co/dacursci/sistemasycomputacion/docs/SystemsEng/SEHandbookv3_2006.pdf.

The path moving up the "right leg" progresses from component integration, to system verifications against the requirements, and, finally, a validation that the delivered system operates correctly in its target environment. Notably—in the context of engineering *trusted autonomous systems*—the key assumption behind the "V" model is that the capabilities that need to be tested on the way up (i.e., on the right leg) are all known and put in place on the way down (i.e., on the left leg). For example, if the final system needs to be simulated at the tail end of the V&V process, then a simulation capability requirement must have been introduced into the pathway (on the way down), built into the system, and then tested (on the way up). There is no provision, in this classic "V" approach to engineering systems, to add or revisit needs or capabilities after specific milestones have been reached. The agile models of the acquisition process (i.e., the software-intensive and software-dominant models 2 and 6, respectively) offer some flexibility by allowing multiple builds of the software component of a system, but are otherwise as inflexible as the classic "V" in the overall system design and performance.

Various approaches to generalizing the classic "V" engineering model (so that it better accommodates the unique requirements and challenges of developing autonomous systems, and includes elements that help build qualities that engender trust into systems as their autonomous capabilities are being designed) are possible. For example, Palmer, at al.,[620] propose a "Trust V" framework that consists of a "toolbox" of reusable and adaptive trust-building techniques, and is designed with a view towards emphasizing the commonality across system architectures. Each technique is intended to be evaluated for applicability and to evolve as the system's autonomous capabilities are specified and developed. Examples of specific techniques include: (1) *semantic Q&A capability*, whereby an developer/operator can query the system with questions such as "Why did the system perform action X?";[621] (2) future *scenario prediction*, in which the developer/operator can query the system with questions such as "What will the system do next?"; (3) *Turing test*, in which part of the validation process consists of comparing a system's performance to that of a

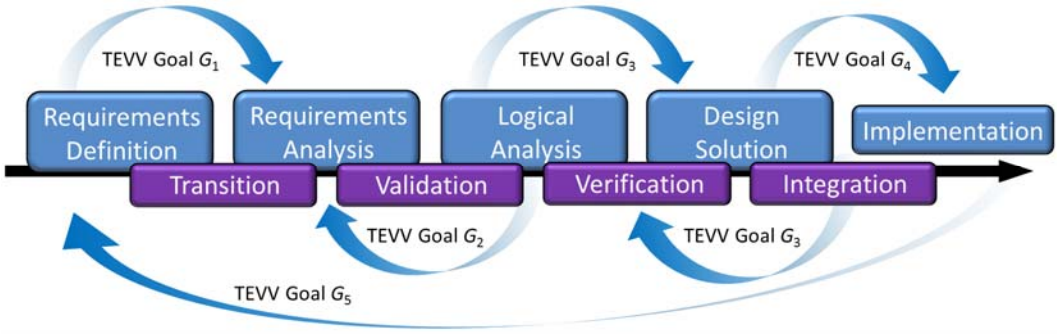
[620] G. Palmer, A. Selwyn, and D. Zwillinger, "The Trust 'V': Building and Measuring Trust in Autonomous Systems," Chapter 4 in *Robust Intelligence and Trust in Autonomous Systems*, edited by R. Mittu, et al., Springer-Verlag, 2016.

[621] Consider IBM Watson's design capability to provide users not only with its "answer" to given question, but with an archive of behind-the-scenes decision that led to that answer along with the confidence the AI system has in each decision. Ref: D. Ferrucci, et.al., "Watson: Beyond Jeopardy!," *Artificial Intelligence* 199/200, 2013.

human expert; and (4) *calibrated trust*, in which an operator's decision making is enhanced by the system providing its insight into its own trustworthiness.[622]

The final report published by the ATEVV/WG recommends that the classic "V" be replaced by a flattened model that explicitly couples system development with V&V, with V&V activities expected to occur both during and between each major development activity (see figure 40).

Figure 40. Concept for an Autonomy TEVV Process Model



After figure 2 in Memo, Assistant SecDef(R&E), Autonomy Test and Evaluation, Verification and Validation Technology Investment Strategy, June 2015.

The concept for ATEVV/WG's Autonomy TEVV Process Model (ATPM) is based on the supposition that for future highly autonomous systems, T&E and V&V activities must be emphasized and distributed throughout the complete acquisition process:

> Testing and evaluating future highly autonomous systems will require an increased emphasis on setting verifiable requirements, developing system models traceable to requirements to guide design activities, and verifying and validating emerging subsystems and products throughout the development process. The final, traditional Development Test (DT) and Operational Test (OT) activities must become a final verification of the complete body of evidence leading to and supporting the documentation of the safety, effectiveness, suitability, and survivability of the system. Essentially, the addition of autonomy will require that much of the effort traditionally reserved for final DT and OT of a new system must be shifted to the left, with

[622] E.M. Roth, "Facilitating 'Calibrated' Trust in Technology of Dynamically Changing 'Trust-Worthiness,'" *Trust in Cyberdomains*, Inst. for Human and Machine Cognition, 2009.

the majority of the T&E activities taking place before the completed system is assembled at test ranges for final system level DT and OT.[623]

In the figure, the arrows denote phases during which five autonomy TEVV Goals (G_1,…,G_5) are meant to be achieved:

- **G_1: Methods to assist in requirements development and analysis**

  Precise, structured standards to automate requirement evaluation for testability, traceability, and de-confliction. Focuses on increasing the fidelity and correctness of autonomous system requirements by developing methods and tools to enable the generation of requirements that are, where possible, mathematically expressible, analyzable, and automatically traceable to different levels (or abstractions) of autonomous system design.

- **G_2: Evidence-Based Design and Implementation**

  Assurance of appropriate decisions w/traceable evidence at every level of design to reduce current T&E burden. Focuses on methods and tools need to be developed at every level of design from architecture definition to modeling abstractions to software generation / hardware fabrication, enabling the compositional verification of the progressive design process, thereby increasing test and evaluation efficiency.

- **G_3: Cumulative Evidence through RDT&E, DT, & OT**

  Progressive sequential modeling, simulation, test and evaluation. Focuses on methods to record, aggregate, leverage, and reuse M&S and T&E results throughout the system's engineering lifecycle; from requirements to model-based designs, to live virtual construction experimentation, to open-range testing.

- **G_4: Run Time Behavior Prediction and Recovery**

  Real-time monitoring, just-in-time prediction and mitigation of undesired decisions and behaviors. Focuses on methods leveraging a run-time architecture must be developed that can provably constrain the
  system to a set of allowable, predictable, and recoverable behaviors,

[623] Memo, Assistant SecDef(R&E), *Autonomy Test and Evaluation, Verification and Validation Technology Investment Strategy*, June 2015, p. 8.

shifting the analysis/test burden to a simpler, more deterministic run-time assurance mechanism.

- **G_5: Assurance Arguments for Autonomous Systems**

  Reusable assurance case based on previous evidence building blocks. "Not only do multiple new TEVV methods need to be employed to enable the fielding of autonomous systems, a new research area needs to be investigated in formally articulating and verifying that the assurance argument itself is valid."

DSB/2016 similarly recommends transforming the conventional model of developmental T&E and operational T&E from "discrete segments of the acquisition cycle to an ongoing evaluation and evolution of the technology and concepts within the operational community,"[624] and advocates for a continuing and pervasive surrounding thread of modeling and simulation activities to support the rapid evolution of autonomous system design and performance (to be integrated throughout the lifecycle of a system, from initial concept to operational test and evaluation, and through operator training). For the concept to be fully successful, operators at all levels must become familiar with and employ T&E techniques, incorporating them within routine training operations.

Lethal Autonomous Weapon Systems

Lethal Autonomous Weapon Systems (LAWS) are weapon systems that, once activated, are able to select and engage targets without human intervention; they are also known as human "out of the loop" autonomous weapon systems. To date, there have been only a few weapon systems that select and engage their own targets. One example is the class of *loitering attack munitions* (LAMs).[625] LAMs are cruise missile-like devices that are launched into a general area and whose mission is to loiter, looking for targets according to pre-programmed targeting criteria (e.g., enemy radars, ships or tanks); once a target is detected, the LAM will fly into the target to destroy it. The only currently operational LAM is the Israel Defense Forces (IDF's) *Harpy*, a "fire-and-forget" anti-radar weapon that flies a general search pattern over a designated area to search for enemy radars, which, if one is found, then dive-bombs

[624] Page 22 in *Summer Study on Autonomy*, DoD, DSB, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016.

[625] Andrea Gilli and Mauro Gilli, "The Diffusion of Drone Warfare? Industrial, Organizational and Infrastructural Constraints: Military Innovations and the Ecosystem Challenge," *Security Studies* 25, 2016: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425750.

into it to destroy it. Examples of experimental LAMs that were not operationally deployed include the low-cost autonomous attack system (LOCAAS),[626] designed to target tanks, and *Tacit Rainbow*, a loitering anti-radar munition.[627] Several adversary nations are known to be developing and research on fully autonomous weapons. Amongst them are China, Germany, India, Israel, Republic of Korea, Russia, and the United Kingdom. Robotic systems with a various degree of autonomy and lethality have already been deployed by the United States, the United Kingdom, Israel, and the Republic of Korea.

While there are weapon systems that automatically sense-and-react to incoming threats such as mortal shells and missiles (e.g., C-RAM,[628] *Phalanx*,[629] and *Mantis*[630]), they are not fully autonomous, and are therefore not technically LAWs as defined above (moreover, such systems are currently confined to defensive functions); rather they are examples of *supervised autonomy* (see page []): they act automatically, not autonomously, and are programmed to execute a small set of actions in predictable environments that entail only a very low risk of incurring civilian harm. Of course, as of this writing, there are no existing autonomous weapons. However, since many of the technologies critical to LAWS (e.g., image processing, image classification, tracking, targeting, weapon trajectory planning, etc.) are currently being developed and continually enhanced—including AI algorithms that, at least in the "narrow" sense, and for non-military applications,[631] already far exceed human performance (e.g., *chess*, *Go*, etc.; see earlier discussion)—it may only be a matter of time before a "critical mass" of technologies is reached, and LAWS will require only the will to be built.

DoD Directive 3000.09 (*Autonomy in Weapon Systems*, issued Nov 2012, and set to expire in 2022) prohibits *lethal* fully autonomous robots.[632] And semi-autonomous robots (e.g., human-in-the-loop control) cannot "select and engage individual targets or specific target groups that have not been previously selected by an authorized

[626] M. Hanlon, "Low-Cost Autonomous Attack System successfully flight tested," *New Atlas,* 4 Nov 2005.

[627] C. Kopp, "Precision guided munitions: Rockwell AGM-130A/B and Northrop AGM-136A Tacit Rainbow, *Air Power Australia*, May 1988.

[628] https://en.wikipedia.org/wiki/Counter_Rocket,_Artillery,_and_Mortar.

[629] https://en.wikipedia.org/wiki/Phalanx_CIWS.

[630] https://en.wikipedia.org/wiki/N%C3%A4chstbereichschutzsystem_MANTIS.

[631] "Toy" (i.e., small-scale versions of) AI military applications are beginning to appear, including aerial dogfights and ground combat. See slides in appendix for details.

[632] DoD Directive 3000.09, Autonomy in Weapon Systems, Nov 2012: http://www.dtic.mil/whs/ directives/corres/pdf/300009p.pdf.

human operator," even in the event that contact with the operator is cut off. Autonomous weapon systems may only be used to apply non-lethal, non-kinetic force, such as some forms of electronic attack, against materiel targets;[633] however, it specifically excludes:

> Cyberspace systems for cyberspace operations; unarmed, unmanned platforms; unguided munitions; munitions manually guided by the operator (e.g., laser- or wire-guided munitions); mines; or unexploded explosive ordnance.[634]

DoD Directive 3000.09 stipulates that autonomous systems:[635] (1) will go through a rigorous review and approval process; (2) will be "designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force," and (3) will be used in accordance with all "applicable domestic and international law, in particular, the *law of war*." (italics added by author; see *The legal dimension* on next page.)

There is a subtle caveat to the directive's otherwise consistent emphasis that target selection and prosecution be conducted under the oversight of a human operator:

> Autonomous or semi-autonomous weapon systems intended to be used in a manner that falls outside the policies in subparagraphs 4.c.(1) through 4.c.(3) must be approved by the Under Secretary of Defense for Policy (USD(P)); the Under Secretary of Defense for

[633] In accordance with DoD Directive 3000.3E, *Policy for Non-Lethal Weapons*, 25 April, 2013.

[634] DoDD 3000.09, pp. 1-2. It has been pointed out that this seemingly well-defined policy distinction of applicability may nonetheless introduce a disconnect into DoD policy with respect to "unarmed, unmanned platforms," since such systems, if they malfunction, may still inflict injury or collateral damage to individuals and property. For example, a malfunctioning automated convoy vehicle may injure a person or cause damage that is similar in its effect to collateral damage from an errant autonomous weapon system. This is but one instance of a slew of ambiguity-ridden and ethics-related issues regarding the use of autonomy, a few of which are discussed later in this report. Ref:J. Caton, *Autonomous Weapon Systems: A Brief Survey of Developmental, Operational, Legal, and Ethical Issues*, U.S. Army War College, Strategic Studies Institute, Carlisle, PA, Dec 2015.

[635] Only the UK and US have issued policies on autonomous weapons systems. The United Kingdom's Ministry of Defence stated in a 2011 Joint Doctrine Note that it "currently has no intention to develop systems that operate without human intervention in the weapon command and control chain, but it is looking to increase levels of automation where this will make systems more effective." Ref: DCDC, "Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems," 30 March 2011.

> Acquisition, Technology, and Logistics (USD(AT&L)); and the CJCS before formal development and again before fielding.[636]

Thus, development of LAWS is not strictly forbidden, and can still occur, provided it is explicitly authorized by appropriate leadership.

Digging a bit deeper on the meaning of "law of war" (italicized in the quote on the middle of the previous page), the reference is to DoD's 1200+ page *Law of War Manual*,[637] that—while not legally binding—represents a comprehensive codification of the proper modes of conduct for all military branches. Section 6.5.9 focuses on autonomy in weapon systems:[638]

> The law of war does not specifically prohibit or restrict the use of autonomy to aid in the operation of weapons. In fact, in many cases, the use of autonomy could enhance the way law of war principles are implemented in military operations. For example, some munitions have homing functions that enable the user to strike military objectives with greater discrimination and less risk of incidental harm. As another example, some munitions have mechanisms to self-deactivate or to self-destruct, which helps reduce the risk they may pose generally to the civilian population or after the munitions have served their military purpose.
>
> Although no law of war rule specifically restricts the use of autonomy in weapon systems, other rules may apply to weapons with autonomous functions. For example, to the extent a weapon system with autonomous functions falls within the definition of a "mine" in the CCW Amended Mines Protocol, it would be regulated as such. In addition, the general rules applicable to all weapons would apply to weapons with autonomous functions. For example, autonomous weapon systems must not be calculated to cause superfluous injury or be inherently indiscriminate.

[636] DoD Directive 3000.09, paragraph 4.d, page 3.

[637] Office of General Counsel, *Department of Defense Law of War Manual*, June 2015: http://www.dod.mil/dodgc/images/law_war_manual15.pdf.

[638] Ibid., "Autonomy in Weapon Systems," p. 329.

The legal dimension

Since there are no existing LAWS, there is also no precedent as to their ethical or legal standing. Tackling the legal dimension first, it is accepted practice that all new technologies of warfare must abide by existing international law, in particular International Humanitarian Law (IHL), which is also referred to as the law of armed conflict.[639] The determination of whether or not a new weapon (including autonomous weapon systems) is in accord with IHL is made by assessing the weapon's foreseeable effects based on its design, and its foreseeable use in normal or expected circumstances. That a weapon cannot be assessed in isolation from its expected method of use derives from the long-standing "Article 36" of the Additional Protocol I to the Geneva Conventions of 1949 (AP I).[640]

Since we obviously do not have the space here to examine the full legal ramifications of the development of LAWS—an excellent summary is provided by Krishnan[641]—we will confine our discussion to outlining just the basic issues. Toward that end, any examination of the lawfulness of LAWS must begin with an aspect of IHL known as "jus in bello" (or, *justice in war*), which focuses on four core principles that define the practices that are allowed and prohibited in war:[642] (1) *military necessity*, (2) *distinction*, (3) *proportionality*, and (4) *unnecessary suffering or humanity*. A fifth principle is also sometimes included: *command responsibility* (which refers to the command chain of liability for illegal acts and the failure to act in the face of foreseeable illegal acts, and therefore lies at the cusp of legality and ethics; see discussion below).

The difficulty with applying *any* of these four principles to LAWS—and what renders the whole legality issue so contentious, leading to a recent effort to pre-emptively ban development of autonomous weapon systems (see below)—is that each requires codifying *qualitative* judgements in software.[643] For example, Article 48 of AP I describes the fundamental rule of distinction as follows:

---

[639] *War & Law*, International Committee of the Red Cross: https://www.icrc.org/en/war-and-law.

[640] M. Schmitt, "Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics," *Harvard National Security Journal: Features Online*, 2013.

[641] A. Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Ashgate, 2009.

[642] D. Stewart, "New Technology and the Law of Armed Conflict: Technological Meteorites and Legal Dinosaurs?" in *U.S. Naval War College International Law Studies 87*, edited by R. Pedrozo and D. Wollschlaeger, U.S. Naval War College, 2011.

[643] Part B in *Autonomous Weapon Systems: Technical, Military, Legal, and Humanitarian Aspects*, Expert Meeting, Geneva, Switzerland, 26-28 March 2014.

> In order to ensure respect for and protection of the civilian population and civilian objects, the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives.[644]

Just as some of the policies in DoD Directive 3000.09 entail "Devil is in the details!" requirements[645] (e.g., the call for T&E to include unanticipated emergent behavior, which is laudable as decreed, but also reveals one of the main "technical challenges" for autonomy; see earlier discussion), to establish, via programming, full compliance with the seemingly innocuous "rule of distinction" is beyond current state-of-the-art AI. Autonomous systems are certainly capable of distinguishing among simple objects in relatively static, uncluttered environments—and military-grade sensor systems can easily detect and recognize pre-defined categories of equipment, such as artillery, tanks, and armored personnel carriers—but the complex reasoning necessary to make qualitative judgements in cluttered environments is currently beyond reach.

Consider the clause in Article 48 of AP I (quoted above) that requires distinguishing between "…civilian objects and military objectives." The precise meaning of this clause (and something that would have to codified in a LAWS) is provided by Article 52(1) of Additional Protocol I to the Geneva Conventions of 1949 (AP I):[646]

> Attacks shall be limited strictly to military objectives. In so far as objects are concerned, military objectives are limited to those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.

"Precision" only serves to complicate matters—at least, so far as codifying legal requirements in autonomous software—since the definition of a military objective is heavily context– (and time–) dependent. While certain objects may meet the required criteria in virtually any combat scenario (e.g., tanks, military aircraft, military bases), AI-like reasoning would be necessary to adjudicate more complex cases. For example,

---

[644] J.-M. Henckaerts and L. Doswald-Beck, *Customary International Humanitarian Law, Volume I: Rules*, Cambridge University Press, 2005.

[645] DoD Directive 3000.09 was issued, in part, in anticipation of future ethical and legal issues associated with autonomous weapons systems.

[646] *Guide to the Legal Review of New Weapons, Means and Methods of Warfare*, ICRC, 2006, page 4: http://www.icrc.org/eng/assets/files/other/icrc_002_0902.pdf.

it is possible to imagine objects that are 'obviously' civilian (prior to or in some early phase of a conflict; e.g., hospitals, schools), but that, at some point—*and for a variety of not simply pre-definable reasons*—become military objectives. Codifying ways in which myriad factors may play out in various contexts well enough for an autonomous system to predictably and consistently make the "right" judgement call (whatever "right" means), poses a significant technical challenge.

The ethical dimension

A complementary set of issues to the legal challenges of autonomy is the ethical dimension; i.e., an examination of the moral principles that may be used to guide the development and deployment of LAWS. At its core, and at the risk of oversimplifying an issue that is far from settled, the challenge of building *ethical* AWS entails essentially the same set of conceptual and technical difficulties as that of autonomy in general; only the focus is different. Namely, the consideration of ethics (in ways that are outlined below) imposes an additional layer of constraints on what an AWS is allowed to do in a given context. All of the same challenges as previously identified for the unconstrained "autonomy problem" remain (e.g., inability to test for all possible contexts, innate unpredictability, propensity for unanticipated "surprises," etc.), but the list must be expanded to include provisions for taking only those actions that are deemed "ethically sound." The central difficulty, of course, is to find ways to endow autonomous machines with the capacity for assessing and responding to moral considerations. The obvious first step is to define what "morality" means.

Since we do not have space in this report to examine all of the theories and approaches to morality (e.g., deontological, or rule-based ethics, consequentialism, natural law, social contract ethics, virtue ethics, etc.),[647] we provide only a brief sketch of issues most relevant to autonomy.

There are two basic approaches to defining and codifying morality:[648] (1) *top-down*, in which a set of rules of behavior are imposed on a moral agent (who may also need to calculate the consequences of various possible courses of action), and (2) *bottom-up*, in which an agent learns to develop a set of "morally correct" behaviors on its own, guided by rewards for "good" behavior as it explores various courses of action in a training environment.

[647] The University of San Diego sponsors an Ethics research site (edited by L. Hinman) that contains a vast collection of ethics-related resources, including basic ethics theory: http://ethics.sandiego.edu/.

[648] P. Lin, G. Bekey, and K. Abney, *Autonomous Military Robotics: Risk, Ethics, and Design*, US Department of Navy, Office of Naval Research, 20 Dec, 2008.

There are two broad categories of bottom-up approaches (recall our earlier discussion of methods of engineering desired behaviors in robotic swarms): (1) *manual tinkering*, in which complex systems are assembled "by hand" out of discrete subsystems, and (2) *emergence*, in which moral values emerge holistically as a desired set of behavioral patterns. Unlike top-down approaches, that derive explicitly from existing moral theory, bottom-up approaches, if they rely on any theory at all, do so only as a way to define *tasks* for learning (and not as a way to define how morality itself is to be learned).

The top-down approach is, for obvious reasons, the naturally preferred method for programmers, since rules are "easy" to turn into algorithms ("easy" is in quotes to remind the reader that, in practice, it is anything but; the "Devil is in the details," as we have stressed repeatedly throughout this report). The challenge is to find ways of codifying the precepts of a specified ethical theory (along with the concomitant meta-challenge of articulating why one particular theory is to be used over another).

Adherents to top-down, rule-based approaches are divided between *deontologists* (i.e., those who insist on a specific set of rules being strictly obeyed, always, even if the consequences are "bad"),[649] and *consequentialists* (i.e., those who grant greater weight to the proposition "the ends justifies the means," and therefore tend to forgive minor transgressions of certain rules if the end result is "good"). The latter approach is by far the more difficult one to apply algorithmically since it is impractical—if not impossible—to be able to calculate the utility of every action (apart from the even more fundamental objection that can be made on the grounds that one can imagine a deception can be deemed as "moral" as a truth, in the event that the consequences are the same).

"Three Rules of Robotics"[650]

The science fiction author, Isaac Asimov, foresaw the need to consider ethical rules of behavior for robots more than 70 years ago when he introduced the "three rules of robotics" (TRoR)—almost literally a text-book example of a top-down, rule-based approach to defining morality—in his short story *Runaround*:[651]

[649] "Deontological" means duty-based, denotes an ethical system built around a system of inflexible rules, and is patterned after Immanuel Kant's categorical imperative. Ref: P. Guyer, et al., Kant's *Groundwork of the Metaphysics of Morals*, Rowman & Littlefield Publishers, 1997.

[650] The discussion in this section is based on pages 29-33 in P. Lin, et al., *Autonomous Military Robotics: Risk, Ethics, and Design*, 2008.

[651] I. Asimov, *Robot Visions*, Roc, 1991.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law

In his stories and novels, Asimov explored the implications and difficulties of his TRoR.[652] For example, the first law was almost immediately seen to be incomplete: as stated, it leaves open the possibility that a robot can inflict harm, so long as that harm is *unintentional*. Hence, if a robot does not know, or cannot predict, that an action on its part will harm a human, it need not obey rule 1. It is also not easy to "fix." For example, while the rule may be rewritten to include an explicit caveat to read: "A robot may do nothing that, to its knowledge, will harm a human being; nor, through inaction, knowingly allow a human being to come to harm," it still leaves open the possibility that someone wishing to *exploit the rule*—the italics are inserted to gently prompt the reader to imagine some latter-day operational equivalent where an adversary wishes to exploit the AI-driven ethics-core of an autonomous weapon system—may divide a task among multiple robots, so that no one robot is able to recognize that its own actions might lead to the harm of a human.

Another difficulty is the ambiguity of risk that a robot must factor into its analysis of a situation. The "through inaction" clause of the first law is particularly problematical in this regard. For example, should robots keep a human from approaching too closely to a cliff over which she can fall (and how is the robot to decide what is "too close")? Are there scenarios in which the robot can, while otherwise acting in strict adherence to the first law, also fail to perform its primary tasks simply because a human repeatedly needs "saving" from harm? In an attempt to fix this problem, Asimov considers a simplified version of the first law to read: "A robot may not harm a human being." But this introduces another (arguably, even worse) problem: it allows a robot that *knows it is capable of preventing harm to a human* to execute an action that will harm the human. For example, the robot may know that the life of a human who, say, wanders into a firing line of an automatic weapon may be spared by canceling a "fire" order to the automatic weapons, but—if the firing sequence is triggered at some time prior to the human coming into range—the robot may fail to prevent the harm because (under the simplified form of rule one) it is no longer strictly required to act.

[652] "Laws of Robotics," *On-line Encyclopedia of Science Fiction*, 5 Nov 2016: http://www.sf-encyclopedia.com/entry/laws_of_robotics.

Asimov later added a zeroth law (so-numbered to imply highest priority):[653] "A robot may not harm all humanity or, through inaction, allow humanity to come to harm." The idea was to allow a robot to harm *individual* humans, but only if in so doing the action prevented an 'existential threat' to all of humanity. But how can a robot determine if such a threat exists?

Other authors have attempted to mend the ambiguities and loopholes inherent in Asimov's TRoR. For example: Dilov[654] proposed a fourth law to prevent possible misunderstandings between what is a robot and what is a human: "A robot must establish its identity as a robot in all cases.", and Clarke[655] has introduced an entire "extended set" of the laws of robotics,[656] though admits that this carefully crafted set still entails serious difficulties (e.g., identification of and consultation with stakeholders and how they are affected, quality assurance, liability for harm resulting from either malfunction or proper use, dispute-resolution, etc.)

The two takeaways from this short discussion are: (1) of the multiple approaches to codifying ethics, deontological, top-down approaches are the preferred method, and (2) even this method may yield "ethics algorithms" that, at best, codify morality rules perfectly but produce harmful consequences, and, at worst, are so fundamentally riddled with ambiguities and subtleties of meaning as to be operationally unusable. Sharkey[657] has argued while a robot may have "…rules of ethics … it won't really care; it will follow a *human designer's idea* of ethics," and therefore by fiat will neither abide by a universal set of standards, nor ever be free of innate bias. Arguments of this form have recently culminated in a call to ban LAWS entirely (see next section).

[653] I. Asimov, *Robots and Empire*, Doubleday, 1985.

[654] D. Lyuben, *The Way of Icarus*, ISBN 954-739-338-3, 1974.

[655] R. Clarke, "Asimov's Laws of Robotics: Implications for Information Technology," *IEEE Computer* (part 1: December 1993, pp. 53–61; part 2: January 1994, pp. 57–66).

[656] (1) a robot may not act unless its actions are subject to the Laws of Robotics (the overriding "meta law"); (2) a robot may not injure humanity, or, through inaction, allow humanity to come to harm; (3) a robot may not injure a human being, or, through inaction, allow a human being to come to harm, unless this would violate a higher-order law; (4) a robot must obey orders given it by human beings, except where such orders would conflict with a higher-order law; a robot must obey orders given it by superordinate robots, except where such orders would conflict with a higher-order Law; (5) a robot must protect the existence of a superordinate robot as long as such protection does not conflict with a higher-order law; a robot must protect its own existence as long as such protection does not conflict with a higher-order law; (6) a robot must perform the duties for which it has been programmed, except where that would conflict with a higher-order law; and (7) a robot may not take any part in the design or manufacture of a robot unless the new robot's actions are subject to the Laws of Robotics.

[657] N. Sharkey, "Autonomous Robots and the Automation of Warfare," *International Humanitarian Law Magazine* 2, 2012.

Perhaps the strongest proponent of the proposition that robots "…can perform *more* ethically than human soldiers are capable of"—and that this reason alone (despite the ambiguities inherent in codifying any system of ethics) renders their continued development an ethical imperative—is Ronald Arkin, Director of Georgia Tech's College of Computing, and a pioneer of behavior-based robotics technologies.[658]

Arkin's point of departure is to use the fledgling but rapidly proliferating technology of driverless cars[659] (e.g., Google's *Waymo*[660] and Tesla's Autopilot[661]) as an example to illustrate how a heretofore human-operator-dominated domain can come to be accepted *despite* being constrained by many of the same ethical and moral concerns that are part of the debate about autonomous systems. Specifically, driverless cars entail two core ethical questions:[662] (1) the classic "trolley problem"[663] (i.e., "How does one decide who lives and who dies in an unavoidable accident?", and (2) addressing the question, "Should the autonomous vehicle always obey the law to the letter?" Though the latter is already being put to the test—a Google car was recently rear-ended as it came to a legal full stop at a stop sign,[664] and a passenger in a Tesla car was killed while the car was in full autopilot[665]—there are (as of this writing) no calls

[658] R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, CRC Press, 2009.

[659] Autonomous vehicle sales are projected to make up between 2-5% of total vehicle sales in the 2020s (and with a lartge price premium), and between 20-40% of of all vehicle sales (with only a moderate price premium) by the 2030s. Ref: T. Litman, "Autonomous Vehicle Implementation Predictions," *Victoria Transport Policy Institute*, 25 Nov, 2016.

[660] https://waymo.com/tech/.

[661] https://www.tesla.com/autopilot.

[662] R. Arkin, "Ethics and Autonomous Systems: Perils and Promises," *Proceedings of the IEEE*, Oct 2016: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7571204.

[663] The "trolley problem" was introduced by J. Thomson ("The Trolley Problem," *The Yale Law Journal*, Vol. 94, No. 6, May, 1985): "Suppose you are the driver of a trolley. The trolley rounds a bend, and there come into view ahead five track workmen, who have been repairing the track. The track goes through a bit of a valley at that point, and the sides are steep, so you must stop the trolley if you are to avoid running the five men down. You step on the brakes, but alas they don't work. Now you suddenly see a spur of track leading off to the right. You can turn the trolley onto it, and thus save the five men on the straight track ahead. Unfortunately, Mrs. Foot has arranged that there is one track workman on that spur of track. He can no more get off the track in time than the five can, so you will kill him if you turn the trolley onto him. Is it morally permissible for you to turn the trolley?" The trolley problem is also examined in a more recent essay: F. Bongiorni, "Give your car a conscience: Why driverless cars need morals," *New Scientist*, 4 Jan 2017.

[664] K. Naughton, "Humans Are Slamming Into Driverless Cars and Exposing a Key Flaw," *Bloomberg Technology*, 17 Dec 2015.

[665] D. Yadron and D. Tynan, "Tesla driver dies in first fatal crash while using autopilot mode," *The Guardian*, 30 June 2016.

to ban driverless cars;[666] indeed, basic issues of liability are being resolved in courtrooms, in a manner entirely analogous to what has been standard practice for human accidents. The reason, according to Arkin, is simply because it is an accepted fact that "humans are the most dangerous things on the road."[667] Therefore, any technology, such as AI-driven cars, that can demonstrably lead to a saving of even one life on the road and otherwise reduce injuries, is seen—or can be argued to be—a moral imperative. Driverless cars cannot get "drunk," are never distracted, and are not subject to road rage, all of which are major contributing factors to highway accidents.

Arkin argues that LAWS can be viewed in a similar light; that is, autonomy (even if it can never be made "perfect") may be viewed as an "ethical imperative" if it helps save noncombatant lives and otherwise reduce injury. Just as driverless cars arguably save lives by compensating (in part) for human weaknesses on the road (as cited above), warfighters are "on occasion prone to poor judgment, carelessness, or even atrocities in their use of force."[668] Arkin cites specific ethics breeches of soldiers and marines deployed in Operation Iraqi Freedom, including:[669] (1) approximate 10% of soldiers report mistreating noncombatants; (2) less than 50% of soldiers agreed that noncombatants should be treated with dignity and respect; (3) less than 50% of soldiers and marines would report a team member for an unethical behavior; (4) only 43% of soldiers (and 30% of marines) agreed they would report a unit member for unnecessarily damaging or destroying private property; and (5) combat experience, particularly losing a team member, was related to an increase in ethical violations.

Robots, however, may be able to perform better than humans under combat conditions (if not now, then in the foreseeable future):[670] they can be designed without emotions that cloud judgment or result in anger and frustration with events taking place on the battlefield; they can assume greater risk on behalf of noncombatants (since they do not need to have the same "instinct" for self-protection as humans); they can process vastly more information from a greater number of data sources than any human; robots are not subject to stress (that may

[666] The National Highway Traffic Safety Administration (NHTSA) has recently issued guidelines (but not *law*) for self-driving cars: https://www.transportation.gov/sites/dot.gov/files/docs/AV%20policy%20guidance%20PDF.pdf.

[667] R. Arkin, "Ethics and Autonomous Systems: Perils and Promises," p. 1780.

[668] Ibid.

[669] *Surgeon General's Office*, Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05-07, Final Report, Nov. 17, 2006.

[670] R. Arkin, *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, Georgia Tech, Technical Report, GIT-GVU-07-11, 2011: http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

cloud judgement); and—when working in teams of combined human soldiers and autonomous systems—robots have the potential to both independently and objectively monitor the ethical behavior of their teammates (thereby reducing human ethical infractions), and more consistently adhere to the "letter of the law" as prescribed by International Humanitarian Law (IHL) and the Rules of Engagement (ROE).

Arkins advocates the use of an "ethical governor" two-step algorithm (albeit admitting that it is premature to speculate whether achieving effective compliance using this method is feasible): *Step 1—go/no-go decision*, in which all sensor-derived data is first used to determine whether an attack is prohibited under IHL and the ROE. If the attack would violate a constraint (e.g., the requirement that a combatant must be distinguished from a noncombatant), it cannot proceed. If no constrains are violated, the attack may proceed only if attacking the target is required under operational orders (the algorithm rests on binary *yes* and *no* answers for this step);[671] and *Step 2—proportionality test*, in which all mission-prescribed and IHL-mandated criteria are statistically weighted in a "utilitarian manner" to determine if the attack "satisfies all ethical constraints and minimizes collateral damage in relation to the military necessity of the target."[672]

Other, even more ambitious approaches—such as to ""match and possibly exceed human intelligence"[673] in engineering IHL-compliant LAWS have also been suggested. For example, the UK Ministry of Defence believes that some form of "true AI" will be required to make autonomous weapons fully comply with the IHL, defining a system with "true AI" as one that has "a similar or greater capacity to think like a human" and distinguishes that intelligence from "complex and clever automated systems."[674] It continues:

> Autonomous systems will, in effect, be self-aware and their response to inputs indistinguishable from, or even superior to, that of a manned aircraft. As such, they must be capable of achieving the same level of situational understanding as a human. . . . As computing and sensor capability increases, it is likely that many systems, using very complex sets of control rules, will appear and be described as

[671] R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, CRC Press, 2009, pp. 183-184.

[672] Ibid., p. 185.

[673] A. Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Ashgate, 2009.

[674] Joint Doctrine Note 2/11 (JDN 2-11), The UK Approach to Unmanned Aircraft Systems, UK Ministry of Defence, The Development, Concepts and Doctrine Centre, 2011: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/33711/20110505JDN_211_UAS_v2U.pdf.

> autonomous systems, but as long as it can be shown that the system logically follows a set of rules or instructions and is not capable of human levels of situational understanding, then they should only be considered to be automated.[675]

Whether such a "true AI" can ever be developed—or is even necessary to achieve "full autonomy" in weapon systems—has been hotly debated for decades, with proponents equally split on the "nay" and 'yay" sides of the proposition.[676] Expert opinions about the future AI vary greatly, with strong disagreements about timescales and about what particular forms "superhuman AIs" might eventually assume.[677] Arguments on the "nay" side date back at least to the 1960s, with arguments by Taube[678] and Lucas[679] that AI is fundamentally incompatible with Kurt Gödel's incompleteness theorem; and to John Searle's "Chinese room" argument against AI, introduced in 1980.[680]

Arguments on the "yay" side include Moravec's[681] and Kurzweil's[682] well-known predictions that machine intelligence will surpass human intelligence by about 2050, and Bostrom's more recent book-length argument that not only is AI inevitable, and will likely arrive much sooner than later, but is something that we must all as a species be on guard against as a clear and imminent danger;[683] a theme that is echoed, and expanded upon by multiple authors, in Awret, et al.[684] Bostrom outlines

[675] Ibid., at 2-3 to 2-4.

[676] C. Schoenick, et al., "Moving Beyond the Turing Test with the Allen AI Science Challenge," *Allen Institute for Artificial Intelligence Science*: http://arxiv.org/pdf/1604.04315v1.pdf

[677] A survey conducted at the 2012 Singularity Summit (an annual conference of the *Machine Intelligence Research Institute*, founded in 2006 at Stanford University) of AI experts found a wide range of predicted dates when AI will be equal to or surpass "human level general intelligence," with a median value of 2040. Ref: S. Armstrong and K. Sotala, "How we're predicting AI, or failing to," in *Beyond AI: Artificial Dreams*, edited by J.Romportl, et al., Pisen: University of West Bohemia, 2013.

[678] M. Taube, *Computers and Common Sense*, Columbia University Press, 1962.

[679] J. R. Lucas, "Minds, machines, and Godel," *Philosophy*, Vol. 36, April-July, 1961.

[680] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences* 3, 1980.

[681] M. Moravec, *Robot: Mere Machine to Transcendent Mind*, Oxford University Press, 2000.

[682] R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, Penguin Books, 2006.

[683] N. Bostrom, *Superintelligence: Paths, Dangers, and Strategies*, 2nd Edition, Oxford University Press, 2016.

[684] U. Awret, B. Appleyard, and D. Chalmers, editors, *The Singularity: Could artificial intelligence really out-think us (and would we want it to)?,* Journal of Consciousness Studies: Imprint Academic, 2016.

five paths that machine AI may follow to achieve human-level general intelligence; and thereafter examines the implications of what he argues becomes inevitable, namely the rapid emergence (i.e., the "AI singularity"[685]) of a super-intelligence that will far exceed human ability to understand or follow.[686] A sentiment that is perhaps best summarized by physicist Stephen Hawking:

> It [artificial intelligence] would take off on its own, and re-design itself at an ever increasing rate, Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.[687]

But, regardless of whether the AI that is embedded in LAWS is "human like" (and displays human-like qualities of emotion and compassion) or "cliché-robotic" (that strictly—and blindly, without regard to any shades-of-gray meaning and interpretations sandwiched between binary-valued extremes)—obeys all IHL, rules-of-war, and ROE), the widespread adoption of such weapons demonstrably raises myriad legal, ethical, and humanitarian concerns; concerns which—depending on what specific set policies and guidelines emerge from the United Nation's recent foray into a public discussion of these issues (see *Movement to ban LAWS* section below)—may potentially impact DoD's own policies and CONOPs with regard to the use of LAWS.[688]

Towards a universal standard of robotic ethics

The Institute of Electrical and Electronics Engineers (IEEE)—the world's largest professional organization for the advancement of technology[689]—has recently

[685] Ibid.

[686] In Ibid., Chalmers (in his introductory essay, "The Singularity: A Philosophical Analysis") cites the basic argument (which he attributes to I. J. Goode, "Speculations Concerning the First Ultraintelligent Machine," *Advances in Computers* 6, 1965): "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion", and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make."

[687] P. Rodgers, "Beware The Robots, Says Hawking," Forbes, 3 Dec 2014: http://www.forbes.com/sites/paulrodgers/2014/12/03/computers-will-destroy-humanity-warns-stephen-hawking/#5edf5e7c3fee.

[688] *Losing Humanity: The Case against Killer Robots*, International Human Rights Clinic, Human Rights Watch, November 2012: https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

[689] https://www.ieee.org/index.html.

published the first version of *Ethically Aligned Design* (IEEE/EAD) with the expressed purpose of encouraging technologists to prioritize ethical considerations in the creation of autonomous and intelligent technologies.[690] It is *not* (nor is meant to be) a definitive examination of ethics, but rather is an interim product that summarizes key issues and actively invites a public discussion of "how these intelligent and autonomous technologies can be aligned to moral values and ethical principles that prioritize human wellbeing."[691]

The 138 page document includes eight sections, each addressing a specific topic related to AI and autonomous systems (AS), and proceeds from three general principles that apply to all AS/AI systems: (1) embody the highest ideals of human rights, (2) prioritize the maximum benefit to humanity and the natural environment, and (3) mitigate risks and negative impacts as AI/AS evolve as socio-technical systems. The document also proposes a three-pronged approach to embedding *values* into AI/AS systems: *Step 1*—identify the norms and values of a specific community affected by AI/AS; *Step 2*—implement the norms and values of that community within AI/AS; and *Step 3*—evaluate the alignment and compatibility of those norms and values between the humans and AI/AS within that community.

Section six of IEEE/EAD examines the ethical issues surrounding the development and use of autonomous weapons systems (AWS). The top-level recommendation is that technical organizations assume a *meaningful human control* of AWS,[692] with audit trails guaranteeing accountability to ensure such control.

---

[690] *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems*, Version 1 – For Public Discussion, IEEE, Dec 2016: http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf. This document is released under the Attribution-Non Commercial version of the Creative Commons license for any organization to adopt or utilize, thereby helping expedite ethical considerations in the creation of autonomous and intelligent technologies.

[691] Details on how to submit feedback to the document are available on-line at: http://standards.ieee.org/develop/indconn/ec/giecaias_guidelines.pdf.

[692] The phrase "meaningful human control" appears throughout government, academic, and policy forums. Recall that "human control" (without the 'meaningful' clause) also appears in DoD Directive 3000.09. A recent document by the United Nations Institute for Disarmament Research (UNIDIR) traces the origin of the phrase "meaningful human control" to UK NGO Article 36 ("Killer Robots: UK Government Policy on Fully Autonomous Weapons", a commentary on the UK Ministry of Defense's 2011 Joint Doctrine Note on "The UK Approach to Unmanned Systems", at www.article36.org/wp-ontent/uploads/2013/04/Policy_Paper1.pdf), and offers a cogent discussion of how an examination of the issues related to "meaningful human control" can fuel broader analyses of the weaponization of increasingly autonomous technologies. Ref: *The Weaponization of Increasingly Autonomous Technologies*, UNIDIR, 2014: http://www.unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf.

The goal is to:

> …ensure that stakeholders are working with sensible and comprehensive shared definitions of concepts relevant in the space of AWS. We recommend designers not only take stands to ensure meaningful human control, but be proactive about providing quality situational awareness through those autonomous or semi-autonomous systems to the humans using those systems. Stakeholders must recognize that the chains of accountability backward, and predictability forward, also include technical aspects such as verification and validation of systems, as well as interpretability and explainability of the automated decision-making, both in the moment and after the fact.

IEEE/EAD lists 11 basic ethics-related issues that the report recommends be addressed for autonomous weapons systems:[693]

1. *Codes of conduct:* professional organization codes of conduct often have significant loopholes, whereby they overlook holding members' works, the artifacts and agents they create, to the same values and standards that the members themselves are held to, to the extent that those works can be.
2. *Definitions:* confusions about definitions regarding important concepts in AI/AS, and AWS stymie more substantive discussions about crucial issues; i.e., it is more important to know how weapons are *controlled by humans* rather than myopically focus on weapon technology per se. (This issue echoes a basic concern expressed in both DSB/2012 and DSB/2016).[694]
3. *Attribution:* AWS are by default amenable to covert and non-attributable use. Such dynamics can easily lead to unaccountable violence and societal havoc.
4. *Accountability:* There are multiple ways in which accountability for AWS's actions can be compromised. Levels of accountability include those for commanders (e.g., "What are the reasonable standards for commanders to utilize AWS?"), and operators (e.g., "What are the levels of understanding required by operators regarding system state, operational context, and situational awareness?")

[693] *Ethically Aligned Design*, IEEE, pp. 68-79.

[694] *The Role of Autonomy in DoD Systems*, DoD Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012; *Summer Study on Autonomy*, DoD, Defense Science Board, Task Force Report, Office of the Under Sec. of Def. for Acquisition, Technology and Logistics, June 2016.

5. *Predictability:* in another echo of concerns expressed in both DSB/2012 and DSB/2016, though raised here in an ethics-related context, unpredictability is cited by IEEE/EAD as an intrinsic property of complex adaptive systems (i.e., AWS). Modeling and simulation cannot, in general, account for all possible dynamic contexts or situational interactions that an AWS will encounter, nor can the interactions with an adversary's systems be anticipated. The inclusion of real-time learning (as discussed earlier) can only compound the problem.
6. *Spectre of a runaway "AWS race":* the widespread adoption, and therefore de facto legitimization of AWS development, may set geopolitically dangerous precedents. The deployment of AWS may create incentives for further use and development of more sophisticated AWS; a cycle that would incentivize faster, increasingly complex, decision-making in critical situations and conflicts, and make it more difficult for humans to participate in decision making.
7. *Bypassing of ethical constraints:* exclusion of human oversight from the battlespace may lead to inadvertent violation of human rights and inadvertent escalation of tensions. "AWS operating without meaningful human control should be prohibited, and as such design decisions regarding human control must be made so that a commander has meaningful human control over direct attacks during the conduct of hostilities."
8. *Consequences of proliferation:* the variety of direct and indirect customers of AWS will lead to a complex and troubling landscape of proliferation and abuse. There is an ethical mandate to consider the foreseeable use of AWS, and the risk for misuse.
9. *Spectre of spontaneous conflicts:* since one of the key advantages of AWS is their ability to make decision vastly faster than humans, when opposing AWS interact with one another, there is the potential for conflicts to arise and escalate at a pace impossible for humans to understand or follow.
10. *Lack of ethical or legal standards:* there are currently no standards regulating the compliance of autonomous and semi-autonomous weapons systems with relevant ethical and legal standards.
11. *Morality:* the highest-level issue for which there is, as yet, no universally agreed upon set of standards, concerns basic morality, starting with, "What does it even mean?", and cutting across questioning the morality of developing, designing, producing, and deploying AWS. IEEE/EAD suggests that it is incumbent on the technology community responsible for building these systems to attain a basic understanding of the ethical and moral boundaries of their work (but recognizes that the means to do so is encumbered by a current lack of standards).

Movement to ban LAWS

The first international organization focused on fostering an ethical debate about, and instituting a band on the development and deployment of autonomous weapons, was the International Committee for Robot Arms Control (ICRAC), organized in 2009 with the mission statement:[695]

> Given the rapid pace of development of military robotics and the pressing dangers that these pose to peace and international security and to civilians in war, we call upon the international community to urgently commence a discussion about an arms control regime to reduce the threat posed by these systems. We propose that this discussion should consider the following: Their potential to lower the threshold of armed conflict; The prohibition of the development, deployment and use of armed autonomous unmanned systems.

ICRAC's mission statement was fulfilled in 2013 when, during the 23rd of the United Nations General Assembly Human Rights Council, a report was issued and debated on the development and deployment of autonomous weapons.[696] The 24 participating states expressed concerns regarding the use of fully autonomous weapons and indicated an interest in continuing discussions.[697] The Convention on Conventional Weapons (CCW) was deemed the appropriate body to deal with autonomous weapons; a proposal that was made official in Nov 2013 when at the meeting of states parties of the CCW it was decided to convene a four-day meeting of experts on the topic of fully autonomous weapons.

[695] J. Altmann, P. Asaro, N. Sharkey and R.Sparrow, founding members, ICRAC: http://icrac.net/2014/05/icrac-celebrates-successful-fulfillment-of-its-2009-mission/.

[696] Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, United Nations, 23rd Session of the General Assembly, 9 April 2013: http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf.

[697] Pakistan, Morocco, Mexico, Argentina, Cuba, Sierra Leone, Switzerland, Algeria, and Egypt raised deep concerns about the future implications of such weapons and argued that these weapons should be discussed through the perspectives of both human rights and international humanitarian law. The European Union, several of its member states, the United States, and Brazil seemed more eager to define the issue in terms of arms control. Though the UK originally expressed the opinion that existing weapons-use rules apply to fully autonomous weapons, and that it does not support an international ban, the UK government later clarified its position by stating that fully autonomous weapons do not meet the requirements of international humanitarian law. Ref: *Fully Autonomous Weapons*, Reaching Critical Will: http://www.reachingcriticalwill.org/resources/fact-sheets/critical-issues/7972-fully-autonomous-weapons#Campaign.

The first such meeting (of experts on LAWS) was held in May of 2014,[698] with three subsequent meetings held 13-17, April 2015; 11-15, April 2016; and 12-16 December 2016, respectively.[699] Apart from continuing the ethical debate on the use of autonomous weapons on the international stage (as of this writing, it is unclear what specific actions will eventually emerge: a set of guidelines, regulations, or a total ban), the most recent meeting formally established an open-ended Group of Governmental Experts (GGE) related to emerging technologies in the area of lethal autonomous weapons systems (LAWS); which is scheduled to meet during two sessions in 2017 (the first session to be held either 24-28, April 2017 or from 21 to 25 August 2017; the second session from 13 to 17 Nov 2017).[700]

A major non-governmental-organization (NGO) sponsored movement to ban LAWS—the *Campaign to Stop Killer Robots* (CSKR)—was launched in London in April 2013, and consists of five international NGOs, a regional NGO network, and four national NGOs that work internationally (including Article 36,[701] Human Rights Watch,[702] and ICRAC[703]). Building on previous experiences to ban landmines, cluster munitions, and blinding lasers, CSKR's goal is to establish a coordinated international effort to ban the development of fully autonomous weapon systems and to address the challenges to international law posed by these weapons. A chronology of CSKR's efforts to date is available on-line.[704]

AI experts weight in

Apart from the U.N.–led and NGO-sponsored movements to ban LAWS, in July 2015, over 1,000 robotics and artificial intelligence researchers signed a 434 word open

[698] *Report of the 2014 informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, held in Geneva, 13–14 November 2014: https://daccess-ods.un.org/TMP/9073427.91557312.html.

[699] Additional information and links to reports summarizing all four CCW meetings on LAWS are available on-line: http://www.unog.ch/80256EE600585943/(httpPages)/3CFCEEEF52D553D5C1257B0300473B77?OpenDocument.

[700] *Final Document of the Fifth Review Conference*, CCW, held 12-16 Dec 2016: http://www.unog.ch/80256EDD006B8954/(httpAssets)/AF11CD8FE21EA45CC12580920053ABE2/$file/CCW_CONF.V_10_23Dec2016_ADV.pdf.

[701] http://www.article36.org/.

[702] http://www.hrw.org/.

[703] http://icrac.net/.

[704] http://www.stopkillerrobots.org/chronology/.

letter[705] calling for a ban on offensive autonomous weapons (with 20K+ signatories as of Dec 2016):[706]

> AI technology has reached a point where the deployment of [autonomous weapons] is—practically if not legally—feasible within years, not decades, and the stakes are high: autonomous weapons have been described as the third revolution in warfare, after gunpowder and nuclear arms.
>
> Many arguments have been made for and against autonomous weapons, for example that replacing human soldiers by machines is good by reducing casualties for the owner but bad by thereby lowering the threshold for going to battle. The key question for humanity today is whether to start a global AI arms race or to prevent it from starting. If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce.
>
> Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control.

The significance of this open letter derives principally from the caliber of people who signed it. Signatories include some the world's leading experts on AI and robotics, and other basic science and technology leaders: *Stuart Russell* (Professor of Computer Science, director of the Center for Intelligent Systems, and co-author of the standard textbook *Artificial Intelligence: a Modern Approach*),[707] *Nils J. Nilsson* (Department of Computer Science, Stanford University), *Barbara J. Grosz* (Professor of Natural Sciences, Harvard University; former president AAAI), *Yann LeCun* (Director of AI Research at Facebook), *Noam Chomsky* (Professor, MIT, inductee in IEEE Intelligent Systems Hall of Fame), *Elon Musk* (SpaceX, Tesla, Solar City), *Frank Wilczek* (Nobel Laureat, Physics), *Demis Hassabis* (Director of Google's *DeepMind* AI project), and *Stephen Hawking* (Director of research at the Department of Applied Mathematics and Theoretical Physics at Cambridge).

[705] http://futureoflife.org/open-letter-autonomous-weapons/#.

[706] http://futureoflife.org/awos-signatories/.

[707] S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd Edition, Pearson, 2009.

Just as the final draft of this report was being completed (Jan 2017), it was reported that the European Parliament committee on Legal Affairs had voted in favor of granting legal status to robots,[708] categorizing them as *electronic persons* (a plenary vote is scheduled for February, 2017):[709]

> ...the most sophisticated autonomous robots could be established as having the status of electronic persons with specific rights and obligations, including that of making good any damage they may cause, and applying electronic personality to cases where robots make smart autonomous decisions or otherwise interact with third parties independently.

[708] "Robot kill switches & legal status: MEPs endorse AI proposal," RT, 12 Jan 2017: https://www.rt.com/viral/373450-robot-kill-switches-status/.

[709] M. Delvaux, Rapporteur, *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*, European Parliament, 2016.

Conclusions

The military is on the cusp of a major technological revolution as it enters the *Robotic Age,*[710] in which warfare is conducted by unmanned and increasingly autonomous weapon systems, operating across all domains (air, sea, undersea, land, space, and cyber), and across the full spectrum of military operations. The question is not *whether* the future of warfare will be filled with autonomous, AI-driven robots, but *when* and in what *form*. However, unlike the last "sea change" during the Cold War (i.e., the so-called "2nd Offset"),[711] when advanced technologies such as precision-strike weapons, stealth aircraft, smart weapons and sensors, and GPS were developed primarily by DoD-sponsored research and development programs, a successful transition into the Robotic Age (spurred on by DoD's recent "Third Offset Strategy" innovation initiative)[712] will depend critically on how well DoD is able to embrace technologies and innovations that are now being developed mostly in the commercial world. And, while the human warfighter is not going away anytime soon, if ever (even as the depth and breadth of autonomy steadily expand), human operators will not suddenly lose control of existing unmanned systems. A telltale sign that DoD has made a "no looking back" cross-over into the Robotic Age will be when human operators can no longer fully understand, or *predict*, how autonomous systems behave—i.e., when, for the first time, a human operator is as stunned by some weapon system's action as 18-time world Go champion Lee SeDol was by a single move of the AI that defeated him.

Opportunities and Challenges

If and when fully AI-driven autonomous systems finally arrive, they will offer a variety of obvious advantages to the warfighter. For example, they will eliminate the

[710] Robert O. Work and Shawn Brimley, *20YY: Preparing for War in the Robotic Age*, Center for a New American Security, Jan 2014.

[711] J. McGrath, "Twenty-First Century Information Warfare and the Third Offset Strategy," *Joint Forces Quarterly*, National Defense University, Issue 82, 3rd Quarter 2016.

[712] C. Hagel, Transcript of Keynote speech delivered at *Reagan National Defense Forum Keynote*, Ronald Reagan Presidential Library, Simi Valley, CA, November 15, 2014.

risk of injury and/or death to the human operator; offer freedom from human limits on workload, fatigue, and stress; and be able to assimilate high-volume data and make "decisions" based on time scales that far exceed human ability. If robotic swarms are added into the mix, entirely new mission spaces potentially open up as well—e.g., wide-area, long-persistence, surveillance; networked, adaptive electronic jamming; and coordinated attack. There are also numerous advantages to using swarms rather than individual robots, including: *efficiency* (if tasks can be decomposed and performed in parallel), *distributed action* (multiple simultaneous cooperative actions can be performed in different places at the same time), and *fault tolerance* (the failure of a single robot within a group does not necessarily imply that a given task cannot be accomplished).

However, the design and development of autonomous systems also entails significant conceptual and technical challenges, including:

- *"Devil is in the details" research hurdles:* Developers of autonomous systems must confront many of the same fundamental problems that the academic and commercial AI and robotic research communities have struggled for decades to "solve." To survive and successfully perform missions, autonomous systems must be able to sense, perceive, detect, identify, classify, plan for, decide on, and respond to a diverse set of threats in complex and uncertain environments. While aspects of all these "problems" have been solved to varying degrees, there is, as yet, no system that fully encompasses all of these features.

- *Complex and uncertain environments:* Autonomous systems must be able to operate in complex—possibly, a priori unknown—environments that possess a large number of potential states that cannot all be pre-specified or be exhaustively examined or tested. Systems must be able to assimilate, respond to, and adapt to dynamic conditions that were not considered during their design. This "scaling" problem—i.e., being able to design systems that are developed and tested in static and structured environments, and then have them perform as required in dynamic and unstructured environments—is highly nontrivial.

- *Emergent behavior:* For an autonomous system to be able to adapt to changing environmental conditions, it must have a built-in capacity to learn, and to do so without human supervision. It may be difficult to predict, and be able to account for *a priori* unanticipated, emergent behavior (a virtual certainty in sufficiently "complex" systems-of-systems dynamical systems).

- *Human-machine interactions/I:* The operational effectiveness of autonomous systems will depend on the dynamic interplay between the human operator and the machine(s) in a given environment, and on how the system responds, in real time, to changing operational objectives, in concert with the human's

own adaptation to dynamic contexts. The innate unpredictability of the human component in human-machine collaborative performance only exacerbates the other challenges identified on this list.

- *Human-machine interactions/II:* The interface between human operators and autonomous systems will likely include a diverse space of tools that include visual, aural, and tactile components. In all cases, there is the challenge of translating human goals into computer instructions (e.g., "solving" a long-standing "AI problem" of natural language processing), as well as that of depicting the machine's "decision space" in a form that is understandable by the human operator (e.g., allowing the operator to answer the question, "Why did the system choose to take action X?").

- *Control:* As autonomous systems increase in complexity, we can expect a commensurate decrease in our ability to both predict and control such systems—i.e., the "spectre of complacency in complexity." As evidenced by the general nature of recent AI breakthroughs, there is a fundamental tradeoff: either the AI can achieve a given performance level (e.g., it can play the game Go as well as, or better than, a human), or humans can be able to understand how its performance is being achieved).

Apart from these innately technical challenges to developing autonomous systems, there are a set of concomitant acquisition challenges, the origin of which is a recent shift in DoD's innovation-related procurement practices. While the U.S. government has always played an important role in fostering AI research (e.g., ARPA, DARPA, NSF, ONR), most key innovations in AI, robotics, and autonomy are now being driven by the *commercial sector*,[713] and at a pace that DoD's relatively plodding stove-piped acquisition process is ill equipped to accommodate: it takes 91 months (7.6 years), on average, from the start of an analysis of alternatives (AoA) study to initial operational capability (IOC).[714] Even information technology programs—under whose rubric most AI-derived acquisitions naturally fall—have averaged 81 months. By way of comparison, note that within roughly this same interval of time, the commercial AI research community has gone from just experimenting with (prototypes of dedicated

[713] The development of most of the UAVs used in Iraq and Afghanistan was driven not by DoD requirements, but rather by commercial research and development. Ref: "Microsoft, Google, Facebook and more are investing in artificial intelligence: What is their plan and who are the other key players?" *TechWorld*, September 29, 2016.

[714] *Policies and Procedures for the Acquisition of Information Technology*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, March 2009.

hardware-assisted) deep learning techniques,[715] to beating the world champion in Go (along with achieving many other major breakthroughs).

Of course, DoD acquisition challenges, particularly for weapons systems that include a heavy coupling between hardware and software, have been known for decades.[716] However, despite numerous attempts by various stakeholders to address these challenges, the generic acquisition process (at least on the traditional institutional level) remains effectively unchanged. Whatever progress has been made in recent years derives more from *workarounds* instituted by DoD to facilitate "rapid acquisition" of systems,[717] than from wholesale changes applied to stove-piped processes of the acquisition process itself. Some recent progress has been made—e.g., the 2009/2011 National Defense Authorization Acts (NDAA/Sec 804), mandated a new IT acquisition process, which, in turn led to multiple Defense Science Board (DSB) Task Force (TF) studies of the acquisition process. Yet, a notable absence in any of these DSB/TF studies is any explicit mention of autonomy.

Complicating the issue still further is a basic dichotomy between DoD's existing directive on autonomy (DoD Directive 3000.09, issued Nov 2012) and current Test and Evaluation (T&E) and Verification and Validation (V&V) practices. Specifically, Directive 3000.09 requires that weapons systems (italics added by author of this report):[718]

- Go through rigorous hardware and software T&E/V&V, "including analysis of *unanticipated emergent behavior* resulting from the effects of complex operational environments on autonomous or semiautonomous systems."
- "Function as anticipated in realistic operational environments against *adaptive adversaries.*"
- "Are sufficiently robust to minimize failures that could lead to *unintended engagements.*"

[715] The first graphics-processor-based unsupervised deep-learning techniques were introduced in 2009: R. Raina, A. Madhavan, and A. Ng, "Large-scale deep unsupervised learning using graphics processors," *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009.

[716] J. Merritt and P. Sprey, "Negative marginal returns in weapons acquisition," *in American Defense Policy*, Third Edition, edited by R. Head and E. Roppe, John Hopkins Univ. Press, 1973.

[717] Examples include: the U.S. Air Force Rapid Capabilities Office, the U.S. Army's Asymmetric Warfare Group and Rapid Capabilities Office, DoD's Strategic Capabilities Office, and, most recently, SecDef Ashton Carter's Defense Innovation Unit Experimental (DIUx). Ref: B. Fitzgerald, A. Sander, J. Parziale, *Future Foundry: A New Strategic Approach to Military-Technical Advantage*, Center for a New American Security, 2016.

[718] Enclosures 2 and 3 of DoD Directive 3000.09 (*Autonomy in Weapon Systems*, Nov 2012) address T&E and V&V issues, and generally review guidelines, respectively.

Directive 3000.09 further requires that T&E/V&V must "assess system performance, capability, reliability, effectiveness, and suitability under realistic conditions, including possible adversary actions, consistent with the *potential consequences of an unintended engagement or loss of control of the system*."

Yet, existing T&E/V&V practices do not make accommodations for any of the italicized parts of these quoted requirements. Among the many reasons why autonomous systems are particularly difficult to test and validate are: (1) *complexity of the state-space* (it is impossible to conduct an exhaustive search of the vast space of possible system "states" for autonomous systems); (2) *complexity of the physical environment* (the behavior of an autonomous system cannot be specified—much less tested and certified—in situ, but must be tested in concert with interaction with a dynamic environment, rendering the space of system inputs/outputs and environmental variables combinatorically intractable); (3) *unpredictability* (to the extent that autonomous systems are inherently complex adaptive systems, novel or unexpected behavior can be expected to arise naturally and unpredictably in certain dynamic situations; existing T&E/V&V practices do not have the requisite fidelity to deal with emergent behavior); and (4) *human operator trust in the machine* (existing T&E/VV&A practice is limited to testing systems in closed, scripted environments, since "trust" is not an innate trait of a system).

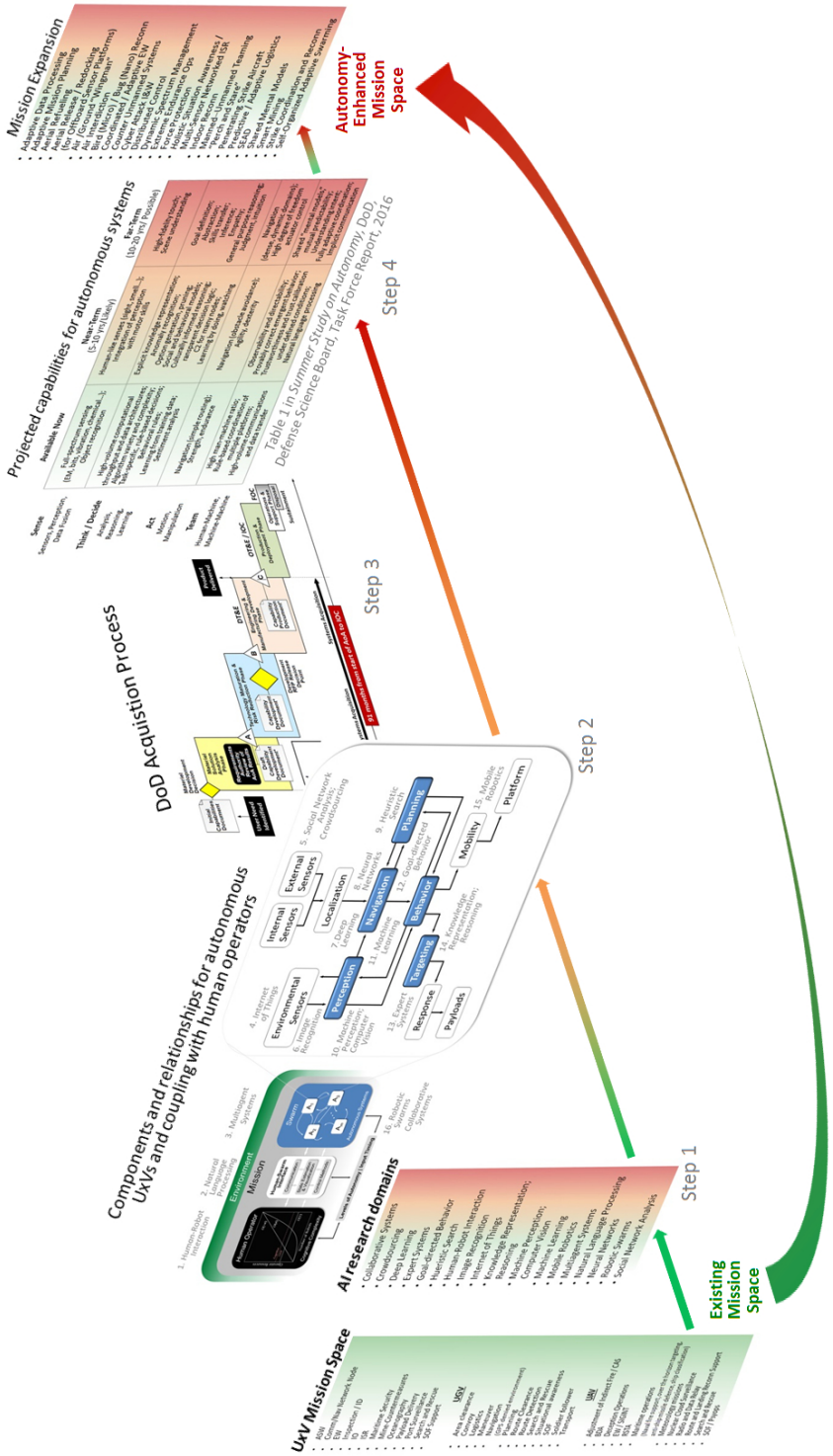
Trust also entails grappling with the issue of *experience* and/or *learning:* to be more effective, autonomous systems may be endowed with the ability to accrue information and learn from experience. But such a capability cannot be certified monolithically, during one "check the box" period of time. Rather, it requires periodic retesting and recertification, the periodicity of which is necessarily a function of the system's history and mission experience. Existing T&E/V&V practices are wholly inadequate to address these issues.

Gestalt of main findings

Figure 41 illustrates, schematically, the key steps involved in extending the existing unmanned systems mission space (e.g., reconnaissance, route clearance, and search and rescue) to one that more fully embraces all that autonomy potentially offers (e.g., self-organized, and self-healing, adaptive swarms). Leaving aside details of the pipeline to the main text, the key (mutually entwined) steps include, starting from bottom of the figure and working our way to the top:

- *Step 1:* Conducting basic AI research across multiple domains (the green-to-red overlay emphasizing that research in different AI areas—e.g., deep learning, image recognition, and robotic swarms—necessarily proceeds at different rates and exists, at any one time, at different levels of maturation).

Figure 41. Key steps in transitioning to new autonomy-enabled mission areas



- *Step 2:* Understanding how individual AI research domains feed into the myriad components that make up autonomous systems, including their coupling with human operators (which further involves the understanding of how human-machine collaborative systems function in specific mission environments).

- *Step 3:* Moving design, development, testing, and accreditation through the DoD acquisition process (and accommodating autonomy's unique set of technical challenges while doing so).

- *Step 4:* Interpreting and projecting the requisite levels of maturity of system capabilities that autonomous systems must possess for specific missions. The autonomous systems that are shown in figure ES-1 are characterized as functions of four broad categories of AI (i.e., *sensing*, *thinking*, *acting*, and *teaming*). Their projected capabilities are indicated as follows: shades of green indicate capabilities that are available now; shades of orange denote near-term capabilities; and increasingly darker shades of red indicate the far-term regime. This table is taken from the DoD's Defense Science Board's most recent study on autonomy,[719] but is intended mostly as a notional place-holder for the kinds of conceptual, technical, and analytical considerations that must be taken into account as the raw capabilities of the autonomous systems that come out of the acquisition process are transformed into new and operationally meaningful missions and missions areas.

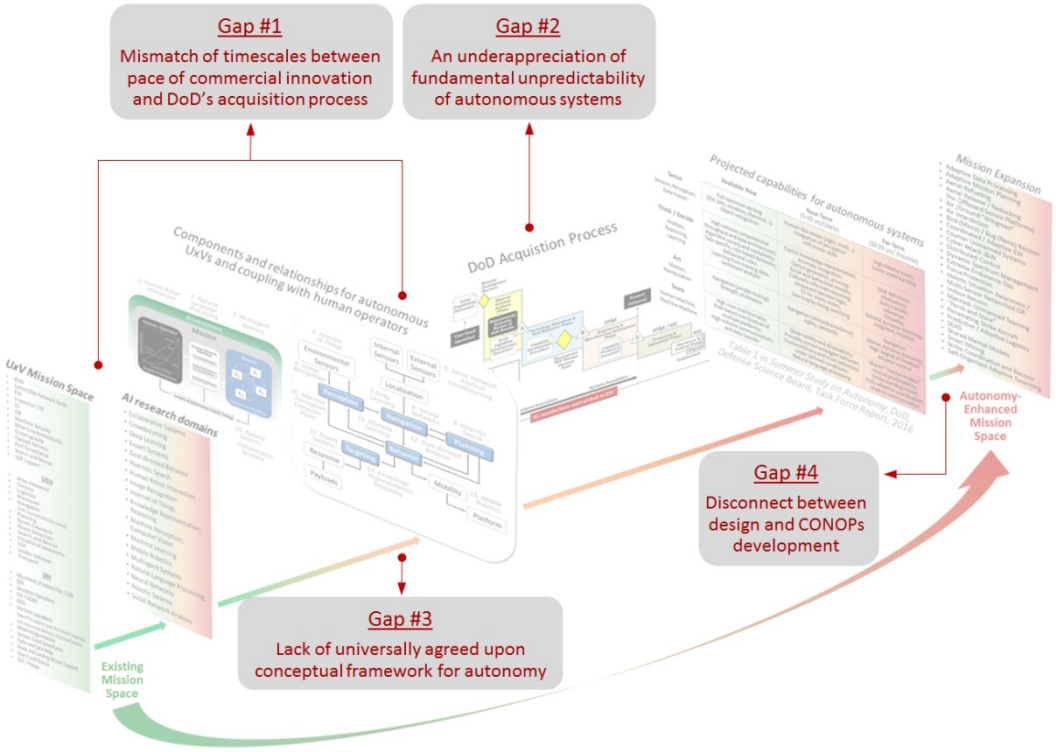
In preparation for DoD's cross-over into the Robotic Age, whenever it arrives, this study has identified four key technical gaps in developing AI-based autonomous systems, wherein opportunities for future analytical studies naturally arise (see figure 42).

These gaps are:

- *Gap 1*: A fundamental mismatch—even *dissonance*—between the accelerating pace (and manner of development and evolution) of technology innovation in commercial and academic research communities, and the timescales and assumptions underlying DoD's existing acquisition process.

[719] Table 1 in *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016: https://www.hsdl.org/?view&did=79464.

Figure 42. Key *gaps* in transitioning to new autonomy-enabled mission areas



- *Gap 2*: An underappreciation of the unpredictable nature of autonomous systems, particularly when operating in dynamic environment, and in concert with other autonomous systems. Existing T&E/V&V practices accommodate neither the basic properties of autonomous systems, as expected by AI and indicated by decades of deep fundamental research into the behavior of complex adaptive systems, nor the requirements they must meet, as weapon systems (as spelled out by DoD Directive 3000.09).

- *Gap 3*: A lack of a universally agreed upon conceptual framework for autonomy that can be used both to anchor theoretical discussions and to serve as a frame-of-reference for understanding how theory, design, implementation, testing, and operations are all interrelated. A similar deficiency exists for understanding the role that trust plays in shaping a human operator's interaction with an autonomous system. The Defense

Science Board's most recent study on autonomy[720] warns that "inappropriate calibration" of trust during "design, development, or operations will lead to misapplication" of autonomous systems, but offers only a tepid definition of trust, and little guidance on how to apply it.

- *Gap 4*: DoD's current acquisition process does not allow for a timely introduction of "mission-ready" AI/autonomy, and there is a general disconnect between system design and the development of concepts of operations (CONOPS). Unmanned systems are typically integrated into operations from a *manned*-centric CONOPS point of view, which is unnecessarily self-limiting by implicitly respecting human performance constraints.

Recommended studies

While not even AI experts can predict how AI will evolve in even the near–term future (much less project its possible course over 10 or more years,[721] or predict AI's impact on the development of military autonomous systems), it is still possible to anticipate many of the key conceptual, technical, and operational challenges that DoD will face in the coming years as it increasingly turns to and more deeply embraces AI-based technologies, and fully enters the "Robotic Age." From an operational analysis standpoint, these challenges can also be used to help shape future studies:

Recommendation 1: ***Help establish dialog between commercial research and development and DoD.***

Institutions specializing in operational analysis are well suited to act as "go betweens" linking the academic and commercial research communities with military culture / operational needs. Assuming that Secretary of Defense

[720] *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016: https://www.hsdl.org/?view&did=79464.

[721] S. Armstrong, K. Sotala, and S. hÉigeartaigh, "The errors, insights and lessons of famous AI predictions – and what they mean for the future," *Journal of Experimental & Theoretical Artificial Intelligence* 26, no. 3, 2014; D. Fagella, "Artificial Intelligence Risk – What Researchers Think is Worth Worrying About," *Tech Emergence,* 20 March 2016: http://techemergence.com/artificial-intelligence-risk/. For the most recent survey of expert opinion see: V. Muller and N. Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in *Fundamental Issues of Artificial Intelligence*, edited by V. Muller, Springer-Verlag, 2016.

Ashton Carter's Defense Innovation Unit-Experimental (DIUx) program survives into the next administration,[722] operationally informed and technically knowledgeable analysts can help stakeholders better "understand" each other. Cross-fertilization with the Naval Postgraduate School (NPS) may also pay dividends.[723]

Recommendation 2: ***Develop an operationally meaningful conceptual framework for autonomy.***

For example, build on lessons learned from the National Institute of Standards and Technology's (NIST's) stalled evolution of its ALFUS (Autonomy Levels for Unmanned Systems) framework, and develop the skeleton of an idea proposed by DoD's Defense Science Board's 2012 report on autonomy.[724]

Recommendation 3: ***Develop measures of effectiveness (MOEs) and measures of performance (MoP) for autonomous systems.***

Develop a methodology by which the effectiveness of autonomous systems can be measured at all levels (e.g., developers, program managers, decision-makers, and warfighters) and across all required functions, missions, and tasks (e.g., coordination, mission tasking, training, survivability, situation awareness, and workload).

Recommendation 4: ***Use nontraditional modeling and simulation (M&S) techniques to help mitigate AI/autonomy-related dimensions of uncertainty.***

As DoD moves into the Robotic Age, M&S is moving away from "simulations as distillations" of real systems (for which M&S has traditionally been used to develop models in order to gain insights into the *real* system), to "simulation-based rules and algorithms as descriptions" of real (i.e., engineered)

[722] DIUx has been established to help facilitate the discovery and development of capabilities and technologies outside DoD's normal acquisition pipeline. Ref: https://www.diux.mil/.

[723] For example: NPS's Consortium for Robotics and Unmanned Systems Education and Research (CRUSER: https://my.nps.edu/web/cruser), and *Autonomous Systems Track* (http://my.nps.edu/web/ast).

[724] *The Role of Autonomy in DoD Systems*, DoD Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

robots and behaviors. It is here, at the cusp between exploring behaviors and prescribing rules that generate them (e.g., engineering *desired* swarm behaviors), that M&S can help mitigate some of the challenges and uncertainties of developing autonomous systems and robotic swarms. For example, while "swarm engineering" methods exist to facilitate the unique design requirements of robotic swarms, no general method exists that maps individual rules to (desired) group behavior.[725]

Multi-agent based modeling techniques[726] are particularly well suited for developing these rules, and, more generally, for studying the kinds of self-organized emergent behaviors expected to arise in coupled autonomous systems (e.g., "How sensitive is an autonomous system's behavior to changes in its physical environment?", "What new command and control architectures will be needed for robotic swarms?", and "How will the control and behavior of a swarm scale with its size and mission complexity?").

Recommendation 5: ***Apply wargaming techniques to help develop new CONOPS.***

Wargaming can be used to help identify and develop new CONOPS, apply lessons-learned from the experience of using deployed systems, explore options to counter uses of autonomy by potential adversaries, and assist in training (e.g., by exploring trust issues in human-machine collaboration). Wargames can also stimulate and nurture a more unified approach to understanding autonomous system performance and behavior, provided that they are conducted with the support and participation from across all military services and domains.

[725] I. Navarro and F. Matia, "An Introduction to Swarm Robotics," *International Scholarly Research Notes*, Vol. 2013, 2013: https://www.hindawi.com/ journals/isrn/2013/608164/.

[726] A. Ilachinski, *Artificial War: Multiagent-Based Simulation of Combat*, World Scientific, 2004. See also: A. Ilachinski, "Modelling insurgent and terrorist networks as self-organized complex adaptive systems," *International Journal of Parallel, Emergent and Distributed Systems* 27, 2012; A. Ilachinski, *AOEWSim: An Agent Based Model for Simulation Interactions Between Off-Board EW Systems and Anti-Ship Missiles*, CNA, DWP-2013-U-004757, 2013; A. Ilachinski and M. Shepko, *FAC/FIAC Simulation (FFSim): User's Guide*, CNA, Annotated Briefing, 2015.

Recommendation 6: ***Develop new T&E/V&V standards and practices appropriate for the unique challenges of accrediting autonomous systems.***

For example, help ameliorate basic gaps in testing in terms of accommodating complexity, uncertainty, and subjective decision environments, by appealing to and exploiting lessons learned from the development and accreditation practices established by the complex system theory and multiagent-based modeling research communities.

Recommendation 7: ***Explore basic human-machine collaboration and interaction issues.***

As autonomy increases, human operators will be concerned less with the manual control of a vehicle, and more with controlling swarms and directing the overall mission: "What are the operator's informational needs (and workload limitations) for controlling multiple autonomous vehicles?" "How do humans keep pace with an accelerating pace of autonomy-driven operations?" "What kinds of command-and-control relationships are best for human-machine collaboration?" "How are human and autonomous-system decision-making practices optimally integrated?" and "What data practices are key to developing shared situation awareness?"

Recommendation 8: ***Explore the challenges of force-integration of increasingly autonomous systems.***

Essentially all force-integration issues are, as yet, undetermined. They must consider not just "low hanging fruit" extensions of existing CONOPS, in which the human component is simply replaced with unmanned systems and "operational value" of human performance is scaled to accommodate "better" performance (e.g., endurance, survivability), but brainstorm heretofore nonexistent tactics, operations, and missions that fully embrace existing and anticipated future autonomous capabilities. What is the tradeoff between large numbers of simple, low-cost (i.e., "disposable") vehicles and small numbers of complex (multi-functional) ones?

The operationalization of robotic swarms, in particular, represents a heretofore largely untapped dimension of the mission space, and will require the development of new CONOPS. The swarm may be used as a radically new form of

precision coordinated "en masse" guided munition; as a self-healing area surveillance network (which includes collecting and assimilating data on an adversary's Internet-of-Things (IoT);[727] or as an adaptive distributed electronic jammer.

Recommendation 9: ***Explore the cyber implications of autonomous systems.***

Explore what new features increased AI-driven autonomy brings to the general risk assessment of increasingly autonomous unmanned systems. On one hand, autonomy may potentially reduce a force's overall vulnerability to jamming or cyber hacking. For example, communications loss over a jammed data link may be compensated for by the ability of autonomous vehicles to continue performing their mission). On the other hand, autonomy itself may also be *more*, not less, vulnerable to a cyber intrusion. For example, an adversary may gain "control," or otherwise deliberately "perturb" the behavior of an autonomous system; it may also be more difficult to detect embedded malware. In the latter context, consider some future variants of incidents such as the Iranian capture of an RQ-170 *Sentinel* in 2011,[728] and the "keylogging" virus that infected the UAV-control-computers at the Creech Air Force Base in Nevada.[729]

Recommendation 10: ***Explore operational implications of ethical concerns over the use of lethal autonomous weapon.***

Analyze issues of accountability, legality, and liability in arguments put forth by various "Ban LAWS" movements. Examine the possible constraints on missions (along with other associated impediments to the design and development of autonomous systems) that may result from an international ban (or set of limits) imposed on the development or deployment of LAWS, such as might come out of the United-Nations-sponsored government experts' negotiations scheduled to take place sometime in 2017.

[727] G. Seffers, "Defense Department Awakens to Internet of Things," *Signal*, 1 Jan 2015: http://www.afcea.org/content/?q=defense-department-awakens-internet-things.

[728] The Iranian government announced that the RQ-170 was captured by its cyber warfare unit: "Iran shows film of captured US drone," BBC News, 8 Dec 2011: http://www.bbc.com/news/world-middle-east-16098562.

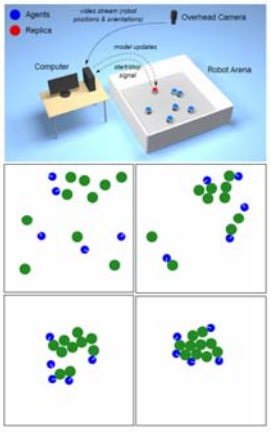
[729] N. Shachtman, "Exclusive: Computer virus hits U.S. drone fleet," *Wired*, 7 Oct 2011.

This page intentionally left blank.

Appendix: recent innovations

The appendix contains selected samples of recent *AI*–, *robot*–, and *swarm*–related research demonstrations and technology innovations, including both commercial and military applications. Though far from complete (the rapid pace of development, alone, renders any static "snapshot" of emerging research and development efforts incomplete, at best, and moot, or dated, or both, at worst), these slides nonetheless offer an additional glimpse into the basic science and engineering that underlies DoD's growing commitment to autonomy.

The examples are all provided in the form of self-contained slides, with additional references to original source material embedded within bulleted summaries. While a few of the examples contained herein are mentioned in the main text (e.g., Google's *AlphaGo* program, "Turing Learning," and robotic self-assembly), most describe work that is not explicitly mentioned elsewhere in this report; and, even in the few cases that were discussed earlier, additional complementary and/or amplifying information is provided.



University of Sheffield, Harvard

- **Turing Learning (TL)** – behavior inference (2016)
  - *Potentially revolutionary* ML-inference method
  - Simultaneously optimizes models and classifiers
    - Two robot swarms: "A" (true) and "B" (learning)
      - ✓ Movements of both "A" and "B" are tracked
    - Two learning systems: *classifier* and *model*
      - ✓ Classifier rewarded for discrimination
      - ✓ Model rewarded for "fooling" the classifier
  - First ever demonstration that an ML-system can infer the behavior of physical robot swarms
    - Moreover, collective behaviors can be directly inferred from motion trajectories of a single agent in the swarm
  - Why revolutionary?

Robots learn how systems behave simply by watching

- Holds true for any observed system, human or machine
- Continued development may radically alter future landscape of real-time adaptive tactics, cyber-related offense/defense capability, vulnerabilities

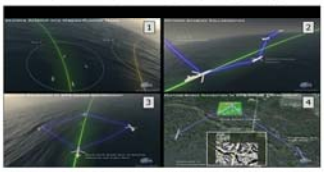
W. Li, M. Gauci, and R. Gross, Turing learning: a metric-free approach to inferring behavior and its application to swarms, Swarm Intelligence, Vol. 10, No. 3, September 2016: http://link.springer.com/article/10.1007%2Fs11721-016-0126-1

http://www.darpa.mil/DDM_Gallery/Small_Gremlins_Web.jpg

- **Gremlins program (DARPA)**
- Goal is to develop innovative technologies and systems enabling aircraft to use C-130 aircraft to launch drone swarms of networked and cooperating unmanned aircraft for EA and reconnaissance missions from standoff ranges, and then recover surviving drones in air when missions are complete
  - Each drone to be between 500 and 1,000 pound, carry 60 pounds of payload and enough fuel to fly out 300 miles, loiter for an hour, and fly back 300 mile
  - Gremlins expected to survive at least 20 missions
  - Target cost of $700,000 per drone
- 2016: Phase 1 contracts awarded (4 companies)
  - "Proof-of-concept" demos to validate air launch and air recovery operations concept of multiple gremlins that can employ intelligence, surveillance and reconnaissance (ISR) and other non-kinetic payloads
- Sample missions
  - A decentralized/distributed, swarm-capable version of the Miniature Air-Launched Decoy Jammer (MALD-J)
  - As prototype "smart" cruise missiles

* Named for imaginary, mischievous imps that became good luck charms of many British pilots during World War II

http://www.darpa.mil/ddm_gallery/code_thumb.jpg



http://www.darpa.mil/DDM_Gallery/CODEPhase2-619-316a.png

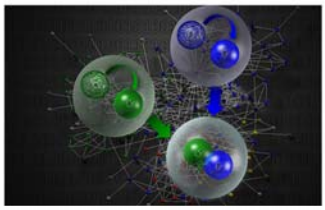
Phase 1: demonstrate potential value of collaborative autonomy in software; 20 autonomous behaviors that would greatly increase mission capabilities of legacy UVs identified (completed Spring/2016)

Phase 2: demonstrate basic autonomous behaviors for two drones (one/Lockheed, other/Raytheon) during a live flight test (announced 6/2016)

Phase 3: one team test their CODE capability using up to six live vehicles cooperating among themselves; plus additional simulated vehicles; one person would command entire swarm (TBD)

• CODE = Collaborative Operations in Denied Environment (DARPA)

- Focus on creating new algorithms and software (for legacy unmanned aircraft) that would extend mission capabilities through autonomy and collaborative behaviors, and improve U.S. forces' ability to conduct operations in denied or contested airspace
  - Software designed to get groups of UAS to work together under a single person's supervisory control
  - UVs would continuously evaluate own states/ environments and recommend coordinated actions to mission supervisor, who would approve/disapprove and direct mission changes
  - CODE-enabled UVs able to find targets and engage them as appropriate under given ROE, leverage nearby CODE-equipped systems with minimal supervision, and adapt to dynamic situations (e.g., emergence of unanticipated threats)
- Goal is to transform UAS operations from requiring multiple operators for each UAS to having a single mission commander simultaneously directing all required UVs
  - Different systems can be mixed/matched with specific capabilities to suit individual missions (instead of depending on a single UAS with integrated capabilities)
  - Designed to increase mission- and cost-effectiveness of legacy assets, reduce development times and costs for future systems, and enable new deployment concepts



http://www.darpa.mil/DDM_Gallery/muse-619-316.jpg

• MUSE = Mining and Understanding Software Enclaves (DARPA)

http://www.darpa.mil/program/mining-and-understanding-software-enclaves

- *Problem:* as computing devices become more pervasive, computing devices have become increasingly more complex
- *Motivation:* data storage and analysis capabilities will be critical for DoD to take advantage of IoT
- *Goal:* to leverage deep program analyses and big data analytics to create a public database containing mined inferences about salient properties, behaviors and vulnerabilities of software
- *Approach:* alter way in which software is built, debugged, verified, maintained and understood; create…
  - Community infrastructure built around a large, diverse and evolving corpus of software drawn from the hundreds of billions of lines of open source code
  - Continuously operational specification mining engine to populate and refine a database containing inferences about useful properties, behaviors and vulnerabilities of the program components in corpus
  - Scalable automated mechanisms to identify and repair code errors
  - Custom programs from existing corpus elements based on properties discovered from data mining activity
- *Multidisciplinary:* programming languages, program analysis, theorem proving and verification, testing, compilers, software engineering, machine learning, databases, statisticians, and multitude of application domains

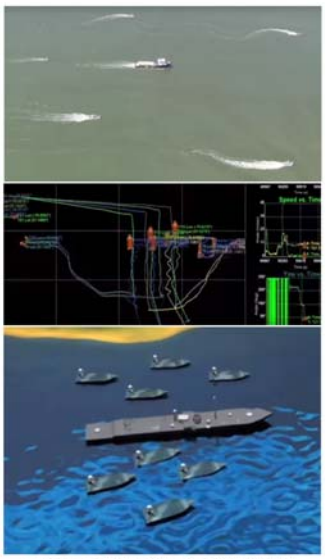
https://3dprintingindustry.com/wp-content/uploads/2016/03/microdrone3d.png

https://www.dvidshub.net/video/504622/perdix-swarm-demo-oct-2016

• 100+ Micro-Drone Swarm Demonstration

(DoD, Strategic Capabilities Office, MIT Lincoln Lab, Oct 2016)

- 103 Perdix drones launched from three F/A-18 Super Hornets at Naval Air Systems Command, China Lake, California
  - Launched from flare dispensers: a flare canister falls after release, small parachute slows fall until canister breaks open, and a drone is released; can also be launched from the ground and sea
  - Each drone transmits to others its location, movement vector
  - Drones do not follow pre-programmed motion scripts
- Micro-drones demonstrated basic swarm behaviors
  - Act autonomously
  - Collective decision-making
  - Adaptive formation flying
  - Self-healing
- Drones are assembled entirely from commercial parts fitted into a 3D-printed fuselage; 30cm long
- Originally created by engineering students at MIT (2013)
- Total project cost: ≈ $20 million
- Ref: DoD Press Release NR-008-17 , 9 Jan 2017: https://www.defense.gov/News/News-Releases/News-Release-View/Article/1044811/department-of-defense-announces-successful-micro-drone-demonstration
- https://www.defense.gov/Portals/1/Documents/pubs/Perdix%20Fact%20Sheet.pdf



http://www.onr.navy.mil/Media-Center/Press-Releases/2014/autonomous-swarm-boat-unmanned-caracas.aspx

• Control Architecture for Robotic Agent Command and Sensing (CARACaS / ONR)

- Platform-agnostic system being developed by ONR; software based on technology from NASA's mars rover
- Can be installed in a variety of small craft to convert them into autonomous (or remote controlled) USVs
  - Sensors consists of a 360-degree electro-optical system with automated targeted recognition platform (CDAS = Contact Detection and Analysis System; a stereo electro-optical infrared (EOIR) system; a radar and automatic identification system (AIS); and a data fusion engine
- August 2014 demo: 13 patrol craft equipped with CARACaS provide a "swarm" escort of an HVU down the James River thalweg
  - The USVs were able to autonomously take station on the escorted vessel with no external inputs using a fused-radar picture
  - When ordered, the swarm broke off escort to surround a contact of interest

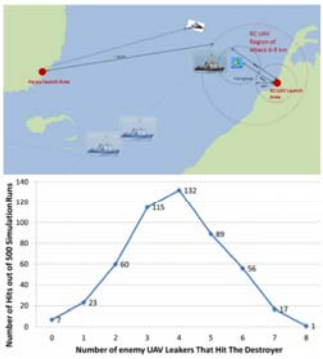
http://spectrum.ieee.org/img/MjgOMzEyMQ-1481915892173.jpeg

– **(CARACaS / ONR): 6 Sep – 3 Oct 2016**
– Builds on 2014 demonstration
  - Boats in in earlier demo did not work together beyond sharing sensor data, as each USV worked alone to determine its own tasking and routes
– New features added between 2014 and 2016
  - *Task allocation:* vessels autonomously assign tasks to each other and work together as a team to perform missions; (Tasks include: patrol, classify, track and trail)
  - *Behavior:* vessels react to situations based on what they perceive through their onboard sensors
  - *Situational awareness:* vessels able to classify unidentified vessels that enter the area as either friend or foe
– Test involved a dynamic "harbor defense" mission
  - Four boats autonomously patrolled area of ≈ 16 sq.nmi
  - React to changing situations, such as an unidentified boat entering the area they were patrolling
  - If an intruding boat was classified as potential enemy, USVs would "decide" which boat would track intruder and which would trail intruder more closely

https://insidedefense.com/daily-news/navy-analyzing-new-missions-autonomous-boats-after-successful-demo

http://www.onr.navy.mil/~/media/Images/220x150/LOCUST%20220x150.ashx; https://youtu.be/AyguXoum3rk

• **Low Cost UAV Swarm Technology (LOCUST) – ONR**
– System to launch swarming UAVs to autonomously overwhelm an adversary
  - Demonstrated in 2015, including the launch of Coyote UAVs ($15K unit cost) capable of carrying varying payloads for different missions
  - Other demos included nine completely autonomous UAV synchronization and formation flight
– Includes tube-based 'cannon' launcher that can send UAVs into the air in rapid succession
– Relies on information-sharing between the UAVs, enabling autonomous collaborative behavior in either defensive or offensive missions
– Small footprint enables swarms of compact UAVs to take off from ships, tactical vehicles, aircraft or other unmanned platforms



- **L. Pham, "UAV swarm attack: protection system alternatives for Destroyers," Thesis, Naval Post Graduate School, 2012***
  - DDG attacked by five to ten drones simultaneously from all directions in conditions of good visibility
    - DDG/Aegis:
      - ✓ Integrated suite of sensors and weapons including jammers, decoys, Standard surface-to-air missiles, a five-inch gun and two Phalanx weapon systems
      - ✓ Augmented with six heavy machine guns on the deck
    - Drones assumed to be made of off-the-shelf components, controlled covertly from nearby fishing vessel
    - Some visually guided, others resemble Israeli "Harpy" loitering drone (with radar guidance)
  - Simulations show that with 8 drones (attacking at 155 mph), average of 2.8 drones get through defenses
    - If DDG equipped with better sensors and more machine guns, at least one drone gets through every time
    - If number of drones > 10, defenses can stop only about 7

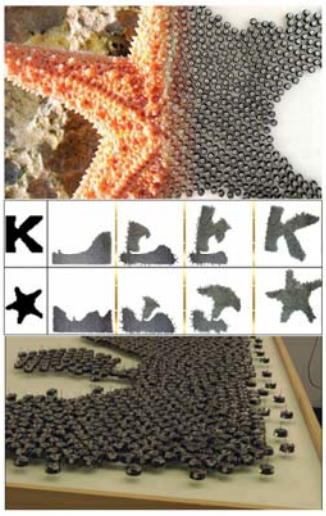
*L. Pham, "UAV swarm attack: protection system alternatives for Destroyers," Thesis, *Naval Post Graduate School*, 2012

"This is an inflection point, this is the first time we've ever had a totally robotic, trans-oceanic-capable ship."
– *DepSecDef*, Robert Work (2016)

- **Navy's "Sea Hunter" Unmanned Surface Vessel Commissioned** (Jan 2016)
  - Developed as part of DARPA's ASW Continuous Trail Unmanned Vehicle (ACTUV) program
  - Top speed = 27 knots; full load displacement = 140 tons (w/40 tons of diesel fuel)
  - All-weather capable; can operate 24/7 (designed for 70 day mission)
  - Has a full autonomy suite, designed to work alongside standard manned craft
    - Maintains operations in line with existing maritime laws (e.g., International Regulations for Preventing Collisions at Sea)
    - 'Sparse' background monitoring allows naval personnel to assume remote control

- ALPHA (University of Cincinnati, 2016)
  - Developed by *Psibernetix* as a research tool for manned and unmanned teaming in a simulation environment
    (*Psibernetix* is contractor to the U.S. Air Force Research Lab)
  - Early versions consistently outperformed other AI opponents (e.g., Air Force Research Lab's in-house flight program)
  - Mature version able to consistently beat (a retired) AF pilot and former battle manager with extensive aerial combat experience
    - Pilot unable to score a single kill against ALPHA on multiple tries
    - Even when ALPHA was put at a severe disadvantage (by limiting its speed, turning, weapons and sensor capabilities), the AI pilot able to beat out other expert human pilots
  - ALPHA able to coordinate a tactical plan in a dynamic combat situation "over 250 times faster than human's can blink"*
  - Runs on a small $35 computer (Raspberry Pi)
    - Trained using $500 commercial PC ("Genetic Fuzzy Tree")

* http://magazine.uc.edu/editors_picks/recent_features/alpha.html



- **Robotic Self-Assembly (Harvard)**
  (Harvard's School of Engineering and Applied Sciences)
  - Introduced a robotic swarm consisting of 1024 small robots – *Kilobots* – that form various shapes
  - Once an initial set of instructions have been delivered, *Kilobots* require no intervention
    - Four robots mark the origin of a coordinate system
    - Other *Kilobots* receive 2-D image to mimic, and use simple "boid-like" behaviors (e.g., follow edge of group, track distance from origin, and maintain relative location)
    - Self-correcting code is included (if problems arise, such traffic jams, nearby *Kilobots* sense the problem and cooperate to fix it)
  - Kilobot hardware / software design is open-source
    - All details are available under a Creative Commons license
  - https://www.eecs.harvard.edu/ssr/projects/progSA/kilobot.html

- **Advanced Robotic Systems Engineering Laboratory (ARSENL)**
  - August, 2015 demo: a swarm of 50 drones controlled by a single operator was successfully launched by a team of students at the Naval Postgraduate School in Monterey, California
    - Launched and flown autonomously in two "sub-swarms" of 25 UAVs each and guided using ARSENL-developed swarm operator interfaces
    - UAVs performed basic leader-follower cooperative behaviors and exchanged information via wireless links
  - Improves on earlier demos using 20 UAVs (May 2015) and 30 UAVs (July 2015)

http://www.sciencealert.com/images/atari-breakout-gif-animated.gif

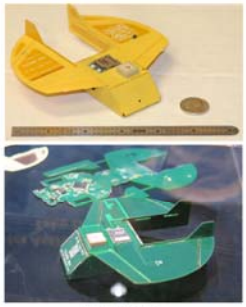
http://cdn.londonandpartners.com/images/explorer-map/tubemap-2012-12.png

- **Google's *DeepMind***
  (http://deepmind.com/research/)
  - 2D games (2015)
    V. Mnih, et al., "Human-level control through deep reinforcement learning," *Nature*, Vol. 518, 26 February 2015
    - Applied Deep Reinforcement Learning (DeepRL) to teach an AI to learn and master – to ***superhuman level*** – 49 different Atari 2600 video games, including *Pong*, *Breakout*, and others
    - First general purpose agent able to continually adapt behavior (across diverse range of scenarios) without any human intervention
  - Navigation (2016)
    A. Graves, et al., "Hybrid computing using a neural network with dynamic external memory," Nature, Vol. 538, 27 October 2016
    - Combines external memory and deep learning to build AI that uses basic reasoning to "self learn" navigation of London underground
  - Encryption (2016)
    M. Abadi, D. Andersen, "Learning to Protect Communications with Adversarial Neural Cryptograph," 21 Oct 2016: https://arxiv.org/abs/1610.06918v1
    - Prototype: neural networks "self taught" themselves a basic encryption technique
    - No a priori cryptographic algorithms are needed

https://upload.wikimedia.org/wikipedia/en/d/d1/Doom_ingame_2.png

- **DeepRL applied to First Person Shooter (FPS)**
  (Carnegie Mellon University, 2016)
  - DeepRL used to teach an AI to play an FPS game (Id Software's* original 1993 version of *Doom*) in a 3D environment
  - First application of DL methods to 3D combat environments
  - Vastly greater "self learning" challenge than 2D games
    - Requires mastery of wide variety of skills: navigating map, collecting items, recognizing / fighting enemies, etc.
  - Note: in 1996, following Gen Krulak's (Commandant of the USMC) directive to use wargames for improving "Military Thinking and Decision Making Exercises," MCCDC developed *Marine Doom*
  - Ref: G. Lample and D. Chaplot, "Playing FPS Games with Deep Reinforcement Learning," https://arxiv.org/abs/1609.05521v1

* http://www.idsoftware.com/



http://defense-update.com/20160419_cicada.html

https://www.mistralsolutions.com/wp-content/uploads/2014/05/cicada-delivery.jpg

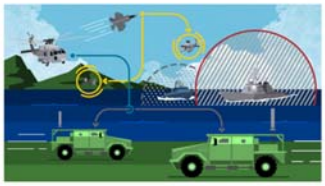
- **Close-In-Autonomous Disposable Aircraft (CICADA)**
  http://www.nrl.navy.mil/tewd/organization/5710/5712/research/CICADA
  - Low-cost, GPS-guided (3D printable) micro disposable air vehicle; essentially a flying circuit board
    - In development since 2006
  - Extremely high packing factor and very low per-unit cost
    - 18 vehicles can be contained in a six-inch cube
  - Can be deployed in large numbers to "seed" an area with miniature electronic payloads
    - Designed to be launched from aircraft (manned or unmanned), balloons, or precision guided munitions, and dispersed in selectable patterns
  - Payloads can be interconnected to form an ad-hoc, self-configuring network
    - Communication nodes and sensors can be placed in a programmable geometric pattern in hostile territory without directly over-flying those regions or exposing human agents on the ground
  - After deployment, CICADA glides to waypoint, enters orbit, and then descends to ground, typically landing within about 15 feet from commanded orbit

DARPA: *recent developments*

http://www.darpa.mil/ddm_gallery/cuavfullsize.jpg

http://www.darpa.mil/DDM_Gallery/hydra-619-316.jpg

- ***Aerial Dragnet* program (2016)**
  DARPA-BAA-16-54 posted 16 Sep
  - Seeks innovative technologies to provide persistent, wide-area surveillance of all UAS operating below 1,000 feet in a large city
  - Focus on protecting military troops operating in urban setting; but has potential to also help protect U.S. metropolitan areas from UAS-enabled terrorist threats
  - Vision is to have a network of surveillance nodes, deploying sensor technologies that look over and between buildings; goal is to establish a continually updated common operational picture (COP) of the airspace at low altitudes
- ***Hydra* program – BAA/2013; 2016 start**
  - Goal is to develop a distributed undersea network of unmanned payloads and platforms to complement manned vessels
  - Will use modular payloads that would provide key capabilities, including ISR and Mine Counter-Measures (MCM)
- ***Spectrum Collaboration Challenge* (SC2)**
  (Registration / Aug 2016)
  - First-of-its-kind collaborative machine-learning (ML) competition to overcome scarcity in the radio frequency (RF) spectrum
  - Multiple teams will compete to build a ML solution for RF scarcity by predicting what other RF devices and potential enemies are doing and figuring out how to best use the available spectrum



http://http://www.bostoncommons.net/wp-content/uploads/2016/01/darpa-airborne-electronic-warfare-ew-777x437.jpg

- ***Adaptive Radar Countermeasures* (ARC)**
  http://www.darpa.mil/program/adaptive-radar-countermeasures
  - Goal is to enable airborne EW systems (AEWs) to automatically generate effective countermeasures against new, unknown and adaptive radars in real-time
  - Vastly different from existing paradigm, in which AEWs must first identify a threat radar to determine the appropriate *preprogrammed* electronic countermeasure (ECM) technique
  - Contract awarded to BAE Systems (Dec 2014): company will deliver a prototype system featuring software algorithms that can detect and counter emerging radar threats
- ***Behavioral Learning for Adaptive Electronic Warfare* (BLADE)**
  http://www.darpa.mil/program/behavioral-learning-for-adaptive-electronic-warfare
  - Goal is to develop the capability to counter new and dynamic wireless communication threats in tactical environments
  - Represents shift from current manual-intensive laboratory-based countermeasure development to an adaptive, in-the-field systems approach using machine learning techniques

• Space Net "AI in the sky" (CIA)

- Essentially a human-crowdsourced machine learning prototype
- In partnership with *Amazon*, *DigitalGlobe*, *Nvidia*, and *CosmiQ Works*
- Goal is to create enormous online database of hi-res images that AIs can use to teach themselves
  - Sustain collaboration between academia, government and industry
- Apply "deep learning" to remote sensing data
  - Automate extraction of features and indicators of activity from satellite imagery
- Testbed: 200,000 curated building footprints across the city of Rio de Janeiro, Brazil (supplied DigitalGlobe)
  - Image dataset publically available on Amazon Web Services: https://aws.amazon.com/public-data-sets/spacenet/

Naval Research Lab: *autonomy*

http://www.nrl.navy.mil/PressReleases/2012/capt-stewart-and-lucas_46-12r_372x248.jpg

Autonomy Research at NRL (< 2012)

- Robotic materials testing system
- *Sea Nimbus* (autonomous distributed sensing and identification technology)
- Shipboard
- Autonomous Firefighting Robot (SAFFiR)
- Autonomous network modeling framework
- Supervisory control of multiple simultaneous autonomous vehicles
- Demonstrated first autonomous boat-to-boat refueling system
- Behavior recognition and threat analysis
- Automatic feature extraction for event detection from videos, sentiment analysis from text documents, and surveillance using deep learning
- Bio-inspired deformable fin UUV
- Developed ISR optimization algorithms for tracking targets in the maritime domain

- Laboratory for Autonomous Systems Research (LASR) – stood up March 2012
  - Provides specialized facilities to support multidisciplinary research in autonomy, human-autonomous system interaction and collaboration, sensor systems, power and energy systems, networking and communications, and platforms
- Research highlights
  - *Flying-Swimmer (Flimmer) UAV/UUV*
    Integrates air delivery with NRL's finned swimmer UUV technologies; goal to achieve fast long-range deployment into hard-to-reach locations
  - *Bio-Inspired UUV for Near-Shore Missions*
  - *Bio-inspired Mobile Autonomous Navy Teams for Information Surveillance and Search*
  - *Ion Tiger Fuel Cell Powered UAV*
  - *Adaptive Testing of Autonomous Systems*
    Uses machine learning to learn minimal number of faults that cause minimal or failed behavior of an autonomous system which is under the control of autonomy software
  - *Swarm Control using Physicomimetics*
    Artificial physics representation in which agents behave as point-mass particles and respond to artificial forces generated by local interactions with nearby particles
  - *Goal-Driven Autonomy (GDA)*
    Focuses on identifying, explaining, and responding to unexpected situations that arise in a complex, dynamic environment
  - *Cognitive Robotics and Human-Robot Interaction*

Naval Post-Graduate School: *autonomy*

https://my.nps.edu/documents/106842137/106977444/ANT_SB.jpg/74539bdd-8f93-4923-9d85-36f201194490?t=1457647229000

https://my.nps.edu/documents/106842137/106977444/ScanEagle1_small.jpg/d2b96ecc-20d1-41a7-ae27-425c97948f23?t=1457642448000

- Center for Autonomous Vehicle Research (CAVR)
  https://my.nps.edu/web/cavr/research
  - Work driven primarily by military needs
  - Ongoing basic research (software & hardware)
    - ANT Glider With Acoustic Vector Sensor
    - ScanEagle Autonomy Extension
    - Helmsman Assist Graphical Interface
    - Collision avoidance
    - Time-Critical Cooperative Path Following
    - Mission management, planning, and execution
    - Field testing of algorithms
    - ISR missions for multiple UAVs (w/UPenn)
- Modeling, Virtual Environments and Simulation (MOVES) Institute
  https://www.movesinstitute.org/
  - Combat-AI simulations
  - Social and Cultural Simulation
  - Cognitive and Perceptual Modeling

http://my.nps.edu/web/cruser/home

- Consortium for Robotics and Unmanned Systems Education and Research (CRUSER)
  - CRUSER Warfare Innovation Continuum (WIC) Workshops Sep 2016: *Developing Autonomy to Strengthen Naval Power*
- Autonomy-related theses – *sampling*
  - Evaluating combined UUV efforts in large-scale MW environment
  - Integrating coordinated path following algorithms to mitigate the loss of communication among multiple UAVs
  - An analysis of the first fifteen years of the DoD framework for Unmanned Ground Systems
  - Human robotic swarm interaction using an artificial physics approach
  - Modular simulation framework for assessing swarm search models
  - Converting a manned LCU into an unmanned surface vehicle (USV): an open systems architecture (OSA) case study
  - Distributed air wings
  - Real-time dynamic model learning and adaptation for UUVs
  - Diver relative UUV navigation for joint human-robot operations
  - A systems engineering analysis of unmanned maritime systems for U.S. Coast Guard missions
  - Effectiveness of unmanned aerial vehicles in helping secure a border characterized by rough terrain and active terrorists
  - Effects of UAV supervisory control of F-18 formation flight performance in a simulator environment
  - Analysis of Nondeterministic Search Patterns for Minimization of UAV Counter-Targeting

ONR: *autonomy science program* (UAS/Code 35)

2016 Topics
- Structured Machine Learning for Scene Understanding
- Understanding Satisficing in Human, Animal, and Engineered Autonomous Systems for Fast Decision Making
- Semantic and Visual Representation of Autonomous System Perceptual Data for Effective Human/Machine Collaboration
- Integrated Autonomy for Log Duration Operations

Other ONR Autonomy-related Programs
- Code 30
  Autonomy of Mixed Initiative Large Teams
  UGV Autonomy / Perception
- Code 31
  Machine Reasoning & Intelligence
- Code 33
  Unmanned Sea Surface Vehicle (USSV)
- Code 34
  Bio-Inspired Autonomous Systems
  Human Robotic Interaction

- Goal = to achieve an affordable heterogeneous mix of naval unmanned systems via unmanned and manned teaming
  - Manifestly interdisciplinary: Deliberately brings together researchers from various traditionally separated domains: air, sea, undersea and ground systems; control theory; computational intelligence; human factors; biology; economics; cognitive science/psychology; and neuroscience
- Focus on four interrelated areas
  - Human collaboration with autonomous systems
  - Perception and intelligent decision making
  - Scalable robust distributed collaboration
  - Intelligence enablers and architectures
- Research Challenges
  - Achieving scalable, survivable, self-organized heterogeneous UxVs with reduced communication requirements
  - Autonomous learning, reasoning, and decision-making in unstructured, contested, dynamic, and uncertain environments
  - Human interaction and collaboration; understanding intent and actions of human team members, adversaries, and bystanders
  - Organic perception and understanding of complex environments

Internet-of-Things (IoT) – 1/2

- IoT = set of IP-addressable devices that interact with environment
  - Typically small, networked devices, designed for simple operation
  - Examples: *thermostats, traffic lights, televisions, mini-drones*
  - Applications: *environmental monitoring, energy infrastructure, transportation*
  - Since possibility for reconfiguration minimal, difficult to mitigate vulnerability
    - Recent study by Hewlett-Packard concluded that 70% of all IoT devices susceptible to compromise
- IoT "universe" is large and growing rapidly
  - Gartner[1] estimates ≈ 6 billion connected things in use during 2016; 25+ billion by 2020
  - Despite these numbers—as a technology—IoT is in its infancy
- Nascent DoD perspective
  - Since "everything," from battlefield uniforms to major weapon systems, are networked…
    …IoT potentially provides vast quantities of real-time data for situational awareness, but presents obvious challenges for cybersecurity, data storage, and analysis
  - Although IoT is on the Defense Information Systems Agency's (DISA's) current strategic "technology watch list"…
    …efforts to define a holistic IoT plan are ad hoc and uncoordinated across services

    "We have people, particularly within our Chief Technology Office, who are looking at what is going on with the Internet of Things, but to begin to say we're trying to define lanes in the road, the answer to that would be no. And quite candidly, it would be premature for us to do that. Industry is still out trying to determine what the Internet of Things consists of."[3]
    – Lt. Gen. Ronnie Hawkins Jr., USAF, DISA director
  - SecDef/Carter embraces IoT as method to "unplug" the military from GPS

[1] http://www.gartner.com/newsroom/id/3165317; [2] http://fortifyprotect.com/HP_IoT_Research_Study.pdf
[3] http://www.afcea.org/content/?q=defense-department-awakens-internet-things

Internet-of-Things (IoT) – 2/2

https://www.praetorian.com/iotmap/#14/30.2679/-97.7440

- Internet of Things Map Project
  (*Praetorian*, Information Security and Risk Consulting firm, Texas: https://www.praetorian.com/)
  - *Goal:* create a searchable database that is the *Shodan*[1] search engine equivalent for Supervisory Control and Data Acquisition (SCADA) devices
  - *Testbed / prototype:* built drone fitted with connected-device tracking appliance and flew it over Austin, Texas
    - Focus on *ZigBee*-enabled smart devices and networks
      ZigBee is a popular smart-home wireless communication standard used by the majority of current IoT devices
    - During 18 minute flight, drone found nearly 1,600 Internet-connected devices
- Vulnerability
  - A recent report by *Cognosec*[2] revealed critical security flaws in *ZigBee* that allows hackers to compromise and take over control of all connected devices on a *ZigBee* network (e.g., door locks, alarm system, light bulbs, etc.)

    https://www.blackhat.com/docs/us-15/materials/us-15-Zillner-ZigBee-Exploited-The-Good-The-Bad-And-The-Ugly-wp.pdf

[1] Shodan is a recently deployed search engine for all Internet-connected devices : https://www.shodan.io/
[2] Cognosec is a European IT security firm – https://www.cognosec.com

Bibliography

1. T. Adams, "Future Warfare and the Decline of Human Decision Making," *Parameters* 31, no. 4, 2001.

2. *Air Force Research Laboratory Autonomy Science and Technology Strategy,* Strategy Report 88ABW-2013-5023, Air Force Research Laboratory, Wright-Patterson AFB, Ohio, Dec 2013.

3. B. Alkire, J. Kallimani, P. Wilson, and L. Moore, *Applications for Navy Unmanned Aircraft Systems*, National Defense Research Institute, RAND Corporation, 2010.

4. *Annual Report to Congress: Military and Security Developments Involving the People's Republic of China 2015*, Office of the Secretary of Defense, 7 April 2015.

5. *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence*, Report of the 2015 Study Panel, Stanford University, Sep 2016: https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf.

6. *Autonomous Horizons: System Autonomy in the Air Force A Path to the Future, Volume I: Human-Autonomy Teaming*, United States Air Force, Office of the Chief Scientist, June 2015.

7. *Autonomous Weapon Systems: Technical, Military, Legal, and Humanitarian Aspects*, Expert Meeting, Geneva, Switzerland, 26-28 March 2014: https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014.

8. *Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume I – Terminology, Version 2.0*, National Institute of Standards and Technology, Special Publication 1011-I-2.0, October 2008.

9. E. Alpaydin, *Introduction to Machine Learning*, Third Edition, MIT Press, 2014.

10. R. Arkin, *Behavior-Based Robotics*, MIT Press, 1998.

11. R. Arkin, *Governing Lethal Behavior in Autonomous Robots*, CRC Press, 2009.

12. R. Arkin, *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, Georgia Tech, Technical Report, GIT-GVU-07-11, 2011: http://www.cc. gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

13. B. Arthur, *Complexity and the Economy*, Oxford University Press, 2015.

14. *Avoiding Surprise in an Era of Global Technology Advances*, Committee on Defense Intelligence Agency Technology Forecasts and Reviews, Division on Engineering and Physical Sciences, National Research Council, National Academy Press, 2005.

15. P. Ball, *Flow*, Oxford University Press, 2009.

16. T. Balch and L. Parker, *editors, Robot Teams*, CRC Press, 2002.

17. M. Barnes, et al., *Designing for Humans in Autonomous Systems: Military Applications*, Army Research Laboratory, ARL-TR-6782, Jan 2014: http://www.arl.army.mil/arlreports/2014/ARL-TR-6782.pdf.

18. P. Bergen and D. Rothenberg, *Drone Wars: Transforming Conflict, Law, and Policy*, Cambridge University Press, 2014.

19. R. Best, *Intelligence technology in the post-cold war era : the role of unmanned aerial vehicles (UAVs)*, Congressional Research Service, the Library of Congress, 1993.

20. M. Bishop, "The Singularity, or How I Learned to Stop Worrying and Love AI," in *Risks of Artificial Intelligence*, edited by V. Muller, Chapman and Hall, pp. 267-280, 2005.

21. C. Blais, *Unmanned Systems Interoperability Standards*, Naval Postgraduate School, NPS-MV-16-001, Sep 2016: http://calhoun.nps.edu/bitstream/handle/10945/50386/NPS-MV-16-001.pdf?sequence=1&isAllowed=y.

22. J. Blom, *Intelligence Technology in the Post-Cold War Era: The Role of Unmanned Aerial Vehicles (UAVs)*, Occasional Paper 37, US Army Combined Arms Center, Combat Studies Institute Press, 2010.

23. M. Boden, "Autonomy: What is it?", *BioSystems* 91 2008.

24. H. Borchert, "Lethal Undersea Drones: The Ultimate Military Game Changer in the Pacific?," *The National Interest*, 9 May 2016.

25. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

26. M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: A review from the swarm engineering perspective," *Swarm Intelligence* 7, no. 1, 2013.

27. S. Brimley, B. Fitzgerald, and K. Sayler, *Game Changers: Disruptive Technology and U.S. Defense Strategy*, Center for a New American Security, Sep 2013.

28. R. Brooks, *Cambrian Intelligence: The Early History of the New AI*, MIT Press, 1999.

29. M. Buckland, *Programming Game AI By Example*, Jones & Bartlett Learning, 2004.

30. R. Bunker, *Terrorist and Insurgent Unmanned Aerial Vehicles: Use, Potentials, and Military Implications*, U.S. Army War College, Strategic Studies Institute, August 2015.

31. J. Caton, *Autonomous Weapon Systems: A Brief Survey of Developmental, Operational, Legal, and Ethical Issues*, U.S. Army War College, Strategic Studies Institute, Carlisle, PA, Dec 2015.

32. J. Cares and J. Dickman, editors, *Operations Research for Unmanned Systems*, Wiley, 2016.

33. M. Chase, et al., *Emerging Trends in China's Development of Unmanned Systems*, RR-990-OSD, RAND Corporation, 2015: http://www.rand.org/pubs/research_reports/RR990.html.

34. J.-L. Chameau, W. Ballhaus, and H. Lin, editors, *Emerging and Readily Available Technologies and National Security: A Framework for Addressing Ethical, Legal, and Societal Issues*, National Academies Press, 2014.

35. M. Clark, K. Kearns, J. Overholt, K. Gross, B. Barthelemy, and C. Reed, *Air Force Research Laboratory Test and Evaluation, Verification and Validation of Autonomous Systems, Challenge Exploration Final Report*, Air Force Research Laboratory, Wright-Patterson AFB, Ohio, Nov 2014.

36. A. Conner-Simons, "System predicts 85 percent of cyber-attacks using input from human experts," *MIT News*, 18 April 2016.

37. G. Coppin and F. Legras, "Autonomy Spectrum and Performance Perception Issues in Swarm Supervisory Control," *Proceedings of the IEEE* 100, no. 3, March 2012: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6108329.

38. L. Davis, M. McNerney, J. Chow, T. Hamilton, S. Harting, and D. Byman, *Armed and Dangerous? UAVs and U.S. Security*, Research Report, RR449, RAND: http://www.rand.org/pubs/research_reports/RR449.html.

39. *Defense Acquisition Handbook*, 16 Sep 2013: https://acc.dau.mil/docs/dag_pdf/dag_complete.pdf.

40. M. Delvaux, Rapporteur, *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics*, European Parliament, 2016.

41. *Department of Defense Developmental Test and Evaluation FY 2015 Annual Report*, Deputy Assistant Secretary of Defense, Developmental Test and Evaluation, March 2016.

42. Department of Defense Instruction 5000.61, *DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)*, USD(AT&L), Dec 9, 2009: http://www.dtic.mil/whs/directives/corres/pdf/500061p.pdf.

43. Department of Defense Instruction 5000.02, *Operation of the Defense Acquisition System*, USD(AT&L), Jan 7, 2015: http://www.navysbir.com/docs/500002p.pdf.

44. M. Dobbing and C. Cole, *Israel and the Drone Wars: Examining Israel's Production, Use and Proliferation of UAVs*, Drone Wars UK, Jan 2014.

45. H. Duan and P. Li, *Bio-inspired Computation in Unmanned Aerial Vehicles*, Springer-Verlag, 2014.

46. G. Dudek, M. Jenkin, and E. Milios, "A taxonomy of multirobot systems." in *Robot Teams*, edited by T. Balch and L. Parker, CRC Press, 2002.

47. A. Dunkelberg, "Laws for L.A.W.S.: Legal Challenges for the Use of Lethal Autonomous Weapons Systems in Times of Armed Conflict," *SSRN*, 29 April, 2016: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2772782.

48. R. Eberhart, Y. Shi, and J. Kennedy, *Swarm Intelligence*, Morgan Kaufmann, 2001.

49. H. Ekbia, *Artificial Dreams: The Quest for Non-Biological Intelligence*, Cambridge University Press, 2008.

50. H. Everett, *Unmanned Systems of World Wars I and II*, MIT Press, 2015.

51. J. Ferber, *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*, Addison-Wesley, 1999.

52. A. Fereidunian, H. Lesani, M. Zamani, K. Sharifi, N. Hassanpour, and S. Mansouri, “A Complex Adaptive System of Systems Approach to Human-Automation Interaction in Smart Grids,” in *Contemporary Issues in Systems Science and Engineering*, edited by M. Zhou, IEEE-Wiley Press, 2015.

53. F. Filbert, “Breaking Integrated Air Defence with Unmanned Aerial Vehicle Swarms,” *JAPCC Journal* 22, Spring-Summer 2016: https://www.japcc.org/breaking-integrated-air-defence-unmanned-aerial-vehicle-swarms/.

54. A. Finn and S. Scheding, *Developments and Challenges for Autonomous Unmanned Vehicles: A Compendium*, Springer-Verlag, 2010.

55. M. Fisher, D. Louise, and M. Webster, "Verifying autonomous systems," *Communications of the ACM* 56, no. 9, 2013.

56. B. Fitzgerald and K. Sayler, *Creative Disruption: Technology, Strategy and the Future of the Global Defense Industry*, Center for a New American Security, June 2014.

57. B. Fitzgerald, A. Sander, and J. Parziale, *Future Foundry: A New Strategic Approach to Military-Technical Advantage*, Center for a New American Security, Dec 2016.

58. D. Gage, “UGV History 101: A Brief History of Unmanned Ground Vehicle (UGV) Development Efforts, *Special Issue on Unmanned Ground Vehicles* 13, no. 3, Summer 1995: http://www.dtic.mil/dtic/tr/fulltext/u2/ a422845.pdf.

59. V. Gazi, B. Fidan, L. Marques, and R. Ordonez, “Robot Swarms: Dynamics and Control,” in *Mobile Robots for Dynamic Environments*, edited by E. Kececi and M. Ceccarelli, Momentum Press, 2015.

60. J. Gertler, *U.S. Unmanned Aerial Systems*, Congressional Research Service, CRS Report for Congress, 3 Jan 2012.

61. D. Gonzales and S. Harting, *Designing Unmanned Systems with Greater Autonomy*, Rand Corporation, 2014.

62. D. Gordon, *Ant Encounters: Interaction Networks and Colony Behavior*, Princeton University Press, 2010.

63. B. Grabowski, *Anticipating the Onset of Autonomy: A Survey of the DoD, Armed Service, and other Federal Agencies’ Outlook on Autonomy*, MITRE Technical Report MP130118, 2013: https://www.mitre.org/sites/default/files/publications/15-1708-anticipating-the-onset-of-autonomy.pdf.

64. B. Grabowski, Big Picture for Autonomy Research in DoD, Keynote presentation at the *Safe and Secure Systems and Software Symposium*, held 09-11 June 2015, Dayton, Ohio, Air Force Research Laboratory: http://www.mys5.org/Proceedings/2015/Day_1/2015-S5-Day1_0805_KEYNOTE_Grabowski.pdf.

65. D. Hambling, *Swarm Troopers*, Archangel Ink, 2015.

66. D. Hambling, "Drone swarms will change the face of modern warfare," *Wired*, 7 Jan, 2016: http://www.wired.co.uk/article/drone-swarms-change-warfare.

67. L. Hardesty, "Making computers explain themselves," *MIT News*, 27 Oct 2016.

68. G. Harrison, *Unmanned Aircraft Systems (UAS): Manufacturing Trends*, Congressional Research Service, January 30, 2013.

69. A. Hassanien and E. Emary, *Swarm Intelligence: Principles, Advances, and Applications*, CRC Press, 2015.

70. M. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, 2003.

71. M. Hati, "Swarm Robotics: A Technological Advancement for Human-Swarm Interaction in Recent Era from Swarm-Intelligence Concept," *International Journal of Science and Research* 5, no. 5, May 2016.

72. S. Hogan, "The Drone Revolution: How Robotic Aviation Will Change the World," *CreateSpace*, 2015.

73. M. Horowitz, "The ethics and morality of robotic warfare: assessing the debate over autonomous weapons," *Daedalus*, Fall 2016.

74. J. Kadtke and L. Wells II, *Policy Challenges of Accelerating Technological Change: Security Policy and Strategy Implications of Parallel Scientific Revolutions*, Center for Technology and National Security Policy, National Defense University, Sep 2014: http://ctnsp.dodlive.mil/files/2014/09/DTP106.pdf.

75. A. Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*, Ashgate, 2009.

76. R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, Penguin Books, 2006.

77. P. Lin, G. Bekey, and K. Abney, *Autonomous Military Robotics: Risk, Ethics, and Design*, US Department of Navy, Office of Naval Research, 20 Dec, 2008: http://ethics.calpoly.edu/onr_report.pdf.

78. S. Lucci and D. Kopec, *Artificial Intelligence in the 21st Century*, Mercury Learning and Information, 2013.

79. G. Marcus, "Moral Machines," *The New Yorker*, 24 Nov, 2012.

80. W. Marra and S. McNeil, "Understanding 'The Loop': Regulating the Next Generation of War Machines," *Harvard Journal of Law & Public Policy* 36, no.3, 2013.

81. M. Matthee, B. Toebes, and M. Brus, editors, Armed Conflict and International Law: In Search of the Human Face, Springer-Verlag, 2013.

82. D. McGlynn, "Robotic Warfare: Should autonomous military weapons be banned?," *CQPress* 25, no. 4, 23 Jan 2015: http://library.cqpress.com/cqresearcher/document.php?id=cqresrre2015012300.

83. Z. Michalewicz and D. Fogel, *How to Solve It: Modern Heuristics*, Springer-Verlag, 2005.

84. J. Miller and S. Page, *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*, Princeton University Press, 2007.

85. I. Millington and J. Funge, *Artificial Intelligence for Games*, 2nd Edition, CRC Press, 2009.

86. R. Mittu, D. Sofge, A. Wagner, and W. Lawless, editors, *Robust Intelligence and Trust in Autonomous Systems*, Springer-Verlag, 2016.

87. M. Moravec, *Robot: Mere Machine to Transcendent Mind*, Oxford University Press, 2000.

88. K. Narasimhan, A. Yala, and R. Barzilay, "Improving Information Extraction by Acquiring External Evidence with Reinforcement Learning," presented at EMNLP 2016, arXiv:1603.07954v3: https://arxiv.org/abs/1603.07954.

89. *NASA Technology Roadmaps, TA 4: Robotics and Autonomous Systems*, National Aeronautics and Space Administration, July 2015: http://www.nasa.gov/sites/default/files/atoms/files/2015_nasa_technology_roadmaps_ta_4_robotics_and_autonomous_systems_final.pdf.

90. *The National Artificial Intelligence Research and Development Strategic Plan*, National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, Oct 2016.

91. I. Navarro and F. Matia, "An Introduction to Swarm Robotics," *International Scholarly Research Notes* 2013, 2013: https://www.hindawi.com/journals/isrn/2013/608164/.

92. I. Navarro and F. Matia, "A Survey of Collective Movement of Mobile Robots," *International Journal of Advanced Robotic Systems* 10, 2013.

93. I. Navarro and F. Matia, "An Introduction to Swarm Robotics," *International Scholarly Research Notices: Robotics* 2013: https://www.hindawi.com/journals/isrn/2013/608164/.

94. J. Nicas, "Criminals, Terrorists Find Uses for Drones, Raising Concerns," *The Wall Street Journal*, Jan. 28, 2015.

95. N. Nilsson, *The Quest for Artificial Intelligence*, Cambridge University Press, 2009.

96. G. Palmer, A. Selwyn, and D. Zwillinger, "The Trust 'V': Building and Measuring Trust in Autonomous Systems," Chapter 4 in *Robust Intelligence and Trust in Autonomous Systems*, edited by R. Mittu, et al., Springer-Verlag, 2016.

97. B. Pendleton and M. Goodrich, "Scalable Human Interaction with Robotic Swarms," *AIAA Infotech@Aerospace (I@A) Conference*, Boston, MA, 2013.

98. Persistent Forecasting of Disruptive Technologies, Committee on Forecasting Future Disruptive Technologies, National Research Council, National Academies Press, 2010: https://www.nap.edu/catalog/12557/persistent-forecasting-of-disruptive-technologies.

99. *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD*, OSD ASDR&E, JASON Report, JSR-16-Task-003, January 2017: https://fas.org/irp/agency/dod/jason/ai-dod.pdf.

100. A. Plaw, M. Fricker, and C. Colon, *The Drone Debate*, Rowman and Littlefield, 2016.

101. D. Poole and A. Mackworth, *Artificial Intelligence*, Cambridge University Press, 2010.

102. R. Proud, J. Hart, and R. Mrozinski, "Methods for Determining the Level of Autonomy to Design into a Human Spaceflight Vehicle: A Function Specific Approach," presented at *Performance Metrics for Intelligent Systems, 16-18 Sep. 2003*, Gaithersburg, MD: https://ntrs.nasa.gov/search.jsp?R=20100017272.

103. S. Rabin, editor, *AI Game Programming Wisdom*, Charles River Media, Inc., 2002.

104. S. Rabin, editor, *AI Game Programming Wisdom 2*, Charles River Media, Inc., 2004.

105. S. Rabin, editor, *AI Game Programming Wisdom 3*, Charles River Media, Inc., 2006.

106. S. Rabin, editor, *AI Game Programming Wisdom 4*, Charles River Media, Inc., 2008.

107. S. Rabin, editor, *Game AI Pro: Collected Wisdom of Game AI Professionals*, CRC Press, 2013.

108. S. Rabin, editor, *Game AI Pro 2: Collected Wisdom of Game AI Professionals*, CRC Press, 2015.

109. T. Rid, *Rise of the Machines: A Cybernetic History*, W. W. Norton and Company, 2016.

110. Robotics, Tele-Robotics, and Autonomous Systems Roadmap, National Aeronautics and Space Administration, Nov 2010: http://www.nasa.gov/pdf/501622main_TA04-Robotics-DRAFT-Nov2010-A.pdf.

111. *Robotics and Autonomous Systems Industry: Final Report*, The Dwight D. Eisenhower School for National Security and Resource Strategy, National Defense University, Spring 2013.

112. *Robotics and Autonomous Systems Industry: Final Report*, The Dwight D. Eisenhower School for National Security and Resource Strategy, National Defense University, Spring 2014.

113. *Robotics and Autonomous Systems Industry: Final Report*, The Dwight D. Eisenhower School for National Security and Resource Strategy, National Defense University, Spring 2015.

114. *Robotics and Autonomous Systems Industry: Final Report*, The Dwight D. Eisenhower School for National Security and Resource Strategy, National Defense University, Spring 2016.

115. *Robotics Strategy White Paper*, Army Capabilities Integration Center - Tank-Automotive Research and Development, Engineering Center Robotics Initiative, 19 March 2009: http://www.dtic.mil/dtic/tr/fulltext/u2/a496734.pdf.

116. M. Rosenberg and J. Markoff, "A.I. Inspiration: The Science Fiction That Frames Discussion," *The New York Times*, 25 Oct 2016.

117. M. Rubenstein, A. Cornejo, and R. Nagpal, "Programmable Self-Assembly in a Thousand-Robot Swarm," *Science* 345, no. 6198, 15 Aug 2014.

118. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Third Edition, Prentice Hall, 2011.

119. S. Russell, S. Hauert, R. Altman, and M. Veloo, "Robotics: Ethics of artificial intelligence," *Nature*, 27 May 2015: http://www.nature.com/news/robotics-ethics-of-artificial-intelligence-1.17611.

120. K. Sayler, P. Scharre and M. Horowitz, *Autonomous Weapons at the UN: A Primer for Delegates*, Center for a New American Security, April 2015.

121. P. Scharre, *Robotics on the Battlefield Part I: Range, Persistence and Daring*, Center for a New American Security, May 2014.

122. P. Scharre, *Robotics on the Battlefield Part II: The Coming Swarm*, Center for a New American Security, Oct 2014.

123. P. Scharre, *Autonomous Weapons and Operational Risk*, Center for a New American Security, Feb 2016.

124. P. Scharre and M. Horowitz, *An Introduction to Autonomy in Weapons Systems*, Working paper, Center for a New American Security, Feb 2015.

125. *Science and Technology (S&T) Priorities for Fiscal Years 2013-2017 Planning*, Memorandum, Secretary of Defense, 19 April 2011: http://www.acq.osd.mil/ chieftechnologist/publications/docs/ OSD%2002073-11.pdf.

126. D. Silver, et al., "Mastering the game of Go with deep neural networks and tree search," *Nature* 529, 28 Jan 2016.

127. J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks* 61, 2015.

128. D. Silver, et al., "Mastering the game of Go with deep neural networks and tree search," *Nature* 529, 28 Jan 2016.

129. L.G. Shattuck, *Transitioning to Autonomy: A human systems integration perspective*, Presentation at Transitioning to Autonomy: Changes in the role of humans in air transportation, March 11, 2015: https://humanfactors.arc.nasa.gov/workshop/autonomy/download/presentations Shaddock%20.pdf.

130. P. W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, Penguin Books, 2009.

131. D. Sloggett, *Drone Warfare: The Development of Unmanned Aerial Conflict*, Skyhorse Publishing, 2014.

132. R. Sparrow, "Robots and respect: Assessing the case against Autonomous Weapon Systems," *Ethics and International Affairs* 30, no. 1, 2016.

133. P. J. Springer, *Military Robots and Drones*, ABS-CLIO, 2013.

134. *Summer Study on Autonomy*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, June 2016.

135. D. Sumpter, *Collective Animal Behavior*, Princeton University Press, 2010.

136. A. Sun, *Cooperative UAV Search and Intercept*, Thesis, Graduate Department of Aerospace Science and Engineering, University of Toronto, 2009.

137. R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

138. P. Swarts, "Air Force looking at autonomous systems to aid war fighters," *Air Force Times*, 17 May 2016.

139. Y. Tan and Z. Zheng, "Research Advance in Swarm Robotics," *Defence Technology*, Vol. 9, No. 1, 2013: http://www.sciencedirect.com/science/article/pii/S221491471300024X.

140. J. Tangney, *Human Systems Roadmap Review*, Briefing Slides, Human Systems Community of Interest, 2016: http://www.defenseinnovationmarketplace.mil/resources/NDIA_Human_Systems_Conference_2016_HSCOI_DistroA_FINAL.pdf.

141. *Technical Assessment: Autonomy*, Department of Defense, Office of Technical Intelligence, Office of the Assistant Secretary of Defense for Research and Engineering, Feb 2015: http://www.defenseinnovationmarketplace.mil/resources/OTI_TechnicalAssessment-AutonomyPublicRelease_vF.pdf.

142. *Technology and Innovation Enablers for Superiority in 2030*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, Oct 2013.

143. *Technology Investment Strategy: 2015-2018*, Autonomy Community of Interest (COI), TEVV Working Group, ASD(R&E) , May 2015.

144. *The Role of Autonomy in DoD Systems*, Department of Defense, Defense Science Board, Task Force Report, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2012.

145. J. Thurnher, "Legal implications of fully autonomous targeting," *JFQ* 67, National Defense University Press, 4th Quarter 2012.

146. V. Trianni, *Evolutionary Swarm Robotics: Evolving Self-Organizing Behaviors in Groups of Autonomous Robots*, Springer-Verlag, 2008.

147. *Unmanned Aerial Systems: DOD Efforts to Adopt Open Systems for Its Unmanned Aircraft Systems Have Progressed Slowly*, GAO-13-651, July 31, 2013: http://www.gao.gov/assets/660/656419.pdf.

148. *Unmanned Aerial Systems: Efforts Made toward Integration into the National Airspace Continue, but Many Actions Still Required*, GAO-15-254T, Dec 10, 2014: http://www.gao.gov/assets/670/667346.pdf.

149. *Unmanned Aerial Systems: Status of Test Sites and International Developments*, GAO- 15-486T, Mar 24, 2015: http://www.gao.gov/assets/670/669214.pdf.

150. *Unmanned Aerial Systems: FAA Continues Progress toward Integration into the National Airspace*, GAO-15-610, Jul 16, 2015: http://www.gao.gov/assets/680/671469.pdf.

151. *Unmanned Systems Integrated Roadmap: FY2011-2036*, U.S. Department of Defense: http://www.acq.osd.mil/sts/docs/Unmanned%20Systems%20Integrated%20Roadmap%20FY2011-2036.pdf.

152. *Unmanned Systems Integrated Roadmap: FY2013-2038*, U.S. Department of Defense: http://archive.defense.gov/pubs/DOD-USRM-2013.pdf.

153. U.S. Government Accountability Office, *Nonproliferation: Agencies Could Improve Information Sharing and End-Use Monitoring on Unmanned Aerial Vehicle Exports*, Washington, D.C., GAO-12-536, July 2012.

154. G. Vasarhelyi, et al., "Outdoor flocking and formation flight with autonomous aerial robots," presented at *IROS 2014 Conference*: arXiv preprint arXiv:1402.3588: http://arxiv.org/abs/1402.3588.

155. D. Vernon, *Artificial Cognitive Systems: A Primer*, MIT Press, 2014.

156. W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, 2014.

157. G. Weiss, editor, *Multiagent Systems*, Second Edition, MIT Press, 2013.

158. R. Whittle, *Predator: The Secret Origins of the Drone Revolution*, Picador, 2015.

159. U. Wilensky and W. Rand, *An Introduction to Agent-Based Modeling*, MIT Press, 2015.

160. S. Wilkerson, C. Korpela, K. Chang, A. Lee, and A. Gadsden, "Aerial Swarms as Asymmetric Threats," presented at *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, June 7-10, 2016. Arlington, VA: http://ieeexplore.ieee.org/document/7502615/.

161. S. Williams, *Arguing A.I.: The Battle for Twenty-first-Century Science*, AtRandom Books, 2002.

162. C. Wills, *Unmanned Combat Air Systems in Future Warfare: Gaining Control of the Air*, Palgrave Macmillan, 2015.

163. J. R. Wilson, *2013 Worldwide UAV Roundup*, American Institute of Aeronautics and Astronautics, July–August 2013.

164. M. Wooldridge, *Introduction to Multiagent Systems*, John Wiley & Sons, 2002.

165. Bob Work, Defense Deputy Secretary, Keynote at the CNAS Inaugural National Security Forum, December 14, 2015: http://www.defense.gov/News/Speeches/Speech-View/Article/634214/cnas-defense-forum.

166. R. O. Work and S., Brimley, *20YY: Preparing for War in the Robotic Age*, Center for a New American Security, Jan 2014.

This page intentionally left blank.

DRM-2017-U-014796-Final

CNA is a not-for-profit research organization
that serves the public interest by providing
in-depth analysis and result-oriented solutions
to help government leaders choose
the best course of action
in setting policy and managing operations.

Nobody gets closer—
to the people, to the data, to the problem.

CNA
ANALYSIS & SOLUTIONS

www.cna.org • 703-824-2000

3003 Washington Boulevard, Arlington, VA 22201