

## Artificial Intelligence: *Emerging Themes, Issues, and Narratives*

Andy Ilachinski

## Abstract

In January 2017, CNA published a 300-plus page report, *AI, Robots, and Swarms*, that examines the conceptual, technical, and operational challenges facing the Department of Defense (DOD) as it pursues AI-based technologies. This white paper is a sequel that brings the 2017 report up to date. It begins with a brief summary of the US Federal Government's and DOD's most recent AI investments, the establishment of the Joint Artificial Intelligence Center (JAIC), and several significant AI-ethics-related events and trends. The rest of the paper is a long narrative that consists of three interwoven parts: Part One compares (and highlights the lack of consensus between) how the academic research community defines AI and how DOD defines it, provides a short history of AI, and offers two complementary views of AI, one as a categorical taxonomy of algorithms, the other as a field of scientific discovery; Part Two summarizes emerging themes and issues, discusses how the AI research community has responded to the COVID-19 pandemic (along with "lessons learned" for DOD), and concludes with evidence that suggests that AI/ML may be entering (or has already entered) an era of diminishing returns; and Part Three introduces a "template of a framework" designed to help bridge the gap between "understanding AI" and operationalizing its military applications. The appendices provide a stand-alone information resource that consists of over 20 high-resolution mindmaps organized around a variety of study-related topics: e.g., taxonomies of AI methods and algorithms; recent breakthroughs and milestones; and gaps, challenges, and limitations of basic AI research. The mindmaps, collectively, contain 800-plus embedded hot-link references.

---

CNA's Occasional Paper series is published by CNA, but the opinions expressed are those of the author(s) and do not necessarily reflect the views of CNA or the Department of the Navy.

## Distribution

Approved for Public Release; Distribution Unlimited.

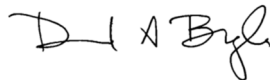
Request additional copies of this document through [inquiries@cna.org](mailto:inquiries@cna.org).

This work was performed under Federal Government Contract No. N00014-16-D-5003.

**Cover image credit:** The infographic was created by the author of this paper.

**Approved by:**

**October 2020**



Dr. David Broyles, Director  
Special Activities and intelligence Program  
Operational Warfighting Division

## Executive Summary

---

We are currently riding a fast-moving artificial intelligence (AI) wave, with reports of “breakthroughs” appearing almost daily. Yet it is important to remember that, since its inception in the 1950s, AI has evolved in fits and starts. A 1958 *New York Times* article (“New Navy Device Learns by Doing”) proclaimed that the Navy had an “embryo” of a “thinking machine” (called a “perceptron,” a precursor of today’s deep neural networks) “that it expects will be able to walk, talk, see, write, reproduce itself, and be conscious of its existence.” These grand expectations were soon dashed, however, and the first of two “AI winters” descended, after fundamental limitations on what perceptrons can achieve in practice were discovered. A second wave of research followed in the 1980s, spurred by more capable neural networks and new methods like reinforcement learning and rule-based expert systems (ES). But after a string of early successes (including “AIs” that defeated a human checkers champion and played backgammon at a human champion performance level), the *second* “AI winter” appeared, this time spanning roughly a decade, from the mid-1990s through 2006, and fueled mainly by, in the case of neural networks, a dearth of sufficiently powerful learning rules and computer power—and, in the case of ES, a lack of sustained progress. A third wave of AI research, launched in 2006—a wave that we are still riding—was spurred by a confluence of three factors: an exponentially dwindling cost of digital storage, coupled with an exponential growth of available data; a new generation of fast learning algorithms for multilayer neural networks; and the exponential growth in computing power, especially that of graphical processing units (GPUs) that speed up learning even more.

A plethora of new methods and “AI successes” have appeared in recent years, none more prominent than AlphaGo, a Go-playing AI developed by Google’s DeepMind, which defeated 18-time world champion Lee Sedol in the game of Go in March 2016. This was a landmark event because the number of possible moves in Go is so vast (well beyond that of chess) that it defies almost any measure of complexity. Indeed, prior to AlphaGo’s victory, most AI experts believed that no AI would defeat a highly ranked human Go player for another 15–20 years. As remarkable as that event was, *less than 20 months later*, an improved learning algorithm (AlphaZero) that required no human-player-data at all and only eight hours of self-play defeated AlphaGo, one hundred games to zero! And, in October 2019, DeepMind’s AlphaStar achieved a grandmaster rating in StarCraft II (which, unlike chess or Go, has  $10^{26}$  actions to choose from at any moment). Demonstrably, the pace of discovery and innovation in AI is accelerating exponentially.

Yet just as demonstrably, such otherwise laudable “successes” mask an Achilles heel that afflicts many of today’s state-of-the-art AIs and portends a larger set of major technical challenges: namely, their “black-box-like” *impenetrability* and *brittleness*. They are “impenetrable” because once they make a “decision,” it is generally hard, if not impossible, to determine the reason(s) *why* they made it. Efforts to develop self-explaining AI systems are underway, but are nascent at best (and are likely to prove as difficult and elusive as “explainable humans”). But how does one trust a system that cannot be understood? AI systems are also generally “brittle” because, even when performing at super-human levels—at games like Go or StarCraft II—they flail when the conditions under which they were trained are changed even slightly (by, say, adding a single new row or column to a Go board). While such “flailing-until-retrained” behavior is of little consequence for game-playing AIs in research labs, it has already proven to be disastrous in real-world settings (e.g., in 2016, a Tesla autopilot that had failed to recognize a white truck against a bright sky—an “environmental exemplar” that was not in the system’s training set—crashed into the truck and killed the driver). A major—thus far, unsolved—challenge for state-of-the-art AI systems is an inherent fragility, or vulnerability to *adversarial attacks*, in which minor changes to images, text documents, or even sound waves (small enough to be *imperceptible to humans*) cause the AI to fail, with possibly catastrophic consequences.

Today’s AI is “*narrow AI*,” not “artificial *general* intelligence” (or AGI) that matches or exceeds human understanding of, and performance on, general intellectual tasks (e.g., the fictional HAL-9000 in *2001: A Space Odyssey*). While true AGI is, at best, decades off (and may, indeed, never materialize), regrettably it is how today’s *narrow AI* is often erroneously perceived. The truth is that today’s state-of-the-art AIs are far from panaceas to general problems.<sup>1</sup> Narrow AI performs well (often, at superhuman levels) only on well-defined, focused tasks and applications (e.g., speech recognition, image classification, and game playing). The more ill-defined and “messy” the problem (think: real world military operations, with all their myriad entwined layers of complexity), the more difficult—and far riskier—the application of state-of-the-art AI and machine learning (ML) techniques.

AI’s technical challenges are not limited to *impenetrability*, *brittleness*, and *adversarial attacks* alone:

- Basic research is plagued with non-reproducibility, in which researchers seeking to reproduce results based on published, peer-reviewed methods are often unable to do so.

---

<sup>1</sup> As further evidence of AI’s seductive power and the omnipresent specter of unwarranted levels of accompanying hype, a 2018 study revealed that, contrary to expectations, in a competition to evaluate the performance of various time series forecasting methods on over 1,000 datasets, the six most accurate methods were *conventional statistical methods*, not “AI” machine learning algorithms.

- AIs generally lack “common sense” knowledge of the world that humans take for granted.
- They generally do not perform well in complex and uncertain environments whose large number of possible states cannot all be pre-specified or exhaustively tested. A recent RAND report documents the major challenges of applying stove-piped verification, validation, and accreditation (VV&A) practices to the development of AI systems.
- They have a limited capacity to transfer “learned abilities” to other problems and domains (and have difficulties integrating prior knowledge).
- AIs do best when their “problem space” is static or slowly changing and struggle with distinguishing causation from correlation, making open-ended inference, and dealing with dynamic environments. The Defense Advanced Research Projects Agency (DARPA) is spearheading research into “lifelong learning” in which AIs learn how to adapt to continuously changing environments, but such work is only at a nascent stage of development.
- AI systems, by their nature, entail emergent behaviors that cannot be anticipated, which may have catastrophic consequences on the battlefield.

Perhaps most egregiously—regarding how “AI successes” are reported to informed but not necessarily technically trained decision makers—AIs falsely seduce us into *expecting* “solutions” to one specific hard problem (such as Carnegie Mellon University’s landmark Pluribus poker-playing AI, which defeated five human professionals in 2019)<sup>2</sup> and scaling naturally to more complex “real world” problems (e.g., military operations). The author witnessed a literal example of AI’s seductive power at a meeting with senior military leaders, during which two AI academic researchers, neither of whom had military operations experience, proffered the “expectation” that a Pluribus-like algorithm can easily be applied to problems in military operations. The reality is that, although the two problem domains are superficially similar (e.g., they both entail decision-making in dynamic environments characterized by imperfect information) and a Pluribus-like approach *may* prove useful (for exploring some operational issues), there are far too many *devil-is-in-the-details*–level issues to expect “quick and easy” results (e.g., defining/acquiring/curating appropriate datasets, developing a sufficiently rich modeling framework to support AI/ML training, and scaling the operational and decision spaces to “real world” dimensions). Although seductively plausible, such scaling efforts are just as likely to languish as “unrealized expectations” than they are to offer genuine insight.

---

<sup>2</sup> Poker is a particularly challenging AI problem because, unlike chess or Go, players can access only partial and/or imperfect information about a game state, and must also be able to deal with bluffing.

Developing a deployable AI system is hard—*very* hard. For example, after IBM’s Watson AI defeated the two highest-ranked human *Jeopardy!* players of all time in a two-game match in 2011,<sup>3</sup> IBM announced a program (that included nearly 50 research partnerships) to develop a version of Watson suitable for dealing with medical problems. Though the project seemed poised to revolutionize medicine, the effort has fallen far short of its lofty goals: as of 2020, most of the partnerships have yet to deliver any commercially viable products, and those products that do exist—say, Watson|*Oncology*—have been criticized for underperformance, such as reportedly recommending “unsafe and incorrect” cancer treatments.

*So, what is AI best suited for?* Today, the answer is *narrow tasks and applications* for which AI has already yielded demonstrably-better-than-human performance, and for which there is ample data. Specific examples include wide-area motion imagery (WAMI); processing, exploitation, and dissemination (PED); and predictive maintenance. The “least suited” uses of AI are those that relegate this technology to black-box “answer machines.” Promises, like the one made in 1958, of far-reaching visions of systems that “walk, talk, see, write, reproduce [themselves], and be conscious of [their] existence”—and even those made in more measured tones of deploying “AI-driven decision aids” on the battlefield—ought to be met with a heavy dose of skepticism. Although it is undeniably seductive, AI is still nascent, laden with both opportunities and major technical hurdles. At the most basic level, AI is a tool—or, better, an *evolving set of tools*—like any other technology that has been used in military operations. It is best used to augment the human decision-making process, not replace it, much as how modeling and simulation tools are still best employed. And the most useful investment strategy to use for developing and deploying specific applications of AI technologies is whichever one is best able to disentangle rigorous science from mal-informed hype.

In January 2017, CNA published a 300-page report, *AI, Robots, and Swarms*, which examines the conceptual, technical, and operational challenges the Department of Defense (DOD) faces as it pursues AI-based technologies.<sup>4</sup> The present occasional paper is intended as a (much shorter) sequel that brings the 2017 report up to date. It begins with a brief summary of the US government’s and DOD’s most recent AI investments, the establishment of the Joint Artificial Intelligence Center (JAIC), and several significant AI ethics–related events and trends. The rest of the paper is a long narrative that consists of three interwoven parts: *Part One* compares (and highlights the lack of consensus between) how the academic research community defines AI and how DOD defines it, provides a short history of AI, and offers two complementary views of AI—one as a categorical taxonomy of algorithms, the other as a field of scientific discovery. *Part Two* summarizes emerging themes and issues, discusses how the

---

<sup>3</sup> A basic research effort (and before applying Watson to medicine) that required *80 man-years* of total development.

<sup>4</sup> Andrew Ilachinski, *AI, Robots, and Swarms: Issues, Questions, and Recommended Studies*, CNA, DRM-20170U-01496, [https://www.cna.org/CNA\\_files/PDF/DRM-2017-U-014796-Final.pdf](https://www.cna.org/CNA_files/PDF/DRM-2017-U-014796-Final.pdf).

AI research community has responded to the COVID-19 pandemic (along with "lessons learned" for DOD), and summarizes recent evidence suggesting that AI/ML may be entering (or has already entered) an era of diminishing returns. *Part Three* introduces a "template of a framework" designed to help bridge the gap between *understanding* AI and *operationalizing* its military applications.

The appendices to this paper provide a stand-alone information resource that consists of more than 20 high-resolution mindmaps organized around a variety of study-related topics (e.g., taxonomies of AI methods and algorithms; recent breakthroughs and milestones; and gaps, challenges, and limitations of basic AI research). When opened in either Adobe Acrobat Reader or Adobe Acrobat Pro, these mindmaps collectively contain an additional 800-plus embedded hot-link references.

The paper concludes with several recommendations for future studies that are listed here in abbreviated form and are, for the most part, self-explanatory. Yet the reader is cautioned that, absent the accompanying narrative, the reasons for these recommendations, along with the many subtleties involved in establishing their contextual significance, will be lost:

- **Recommendation #1:** *Move away from myopically short-sighted "simple" static definitions of AI towards active engagement—and continual re-engagement—of all stakeholders to adapt, adopt, and reconceptualize AI (as new technologies and methods are inevitably introduced) as a holistic Sense → Think → Learn → Act evolutionary cyclic process.*
- **Recommendation #2:** *Embrace the irreducible reality of the inherent challenges associated with developing and deploying AI systems, and develop a set of formal practices and procedures for mitigating fundamental challenges.*
- **Recommendation #3:** *Anticipate, and thus mitigate the otherwise bleak consequences of ignoring, the possibility that deep learning may already have entered (or is soon to enter) an era of diminishing returns, in which ever-increasing computational resources and/or refinements in algorithmic technique yield relatively marginal gains in performance.*
- **Recommendation #4:** *Develop a framework to better enable, foster, and nurture collaborative engagements—and a mutual dialectic—among AI researchers, technology developers, military policy makers, S&T portfolio managers, and warfighters.*
- **Recommendation #5:** *Leverage AI's innate superhuman capacity to search through vast, complex, multidimensional abstract spaces to help human stakeholders discover innovative approaches to "old" problems, and/or discover heretofore unknown solutions to problems not yet recognized.*

[This page intentionally left blank]



# Contents

---

<b>Introduction</b> .....	<b>1</b>
Organization of this paper .....	3
<b>Background</b> .....	<b>5</b>
AI investments .....	5
Joint Artificial Intelligence Center (JAIC).....	5
Defense Advanced Research Projects Agency (DARPA) .....	9
AI ethics .....	10
<b>What Is AI?</b> .....	<b>15</b>
Definitions .....	15
AI research community definitions of AI: <i>little consensus</i> .....	16
DOD definitions of AI: <i>even less consensus</i> .....	17
So what is AI, <i>really?</i> .....	20
Short history.....	21
AI as a <i>categorical taxonomy of algorithms</i> .....	29
AI as a <i>field of scientific discovery</i> .....	32
AI’s perennially persistent fundamental gaps, challenges, and limitations.....	39
<b>Emerging AI Themes and Issues</b> .....	<b>43</b>
Recent trends .....	43
AI “hits” during 2017–2020 .....	48
AI “misses” (2017–2020) .....	53
<i>New AI “challenges” (2017–2020)</i> .....	55
Growing “pushback” against AI/ML .....	56
Unceasing AI “hype” during 2017–2020.....	57
COVID-19 and AI: <i>lessons learned for DOD?</i> .....	58
Specter of a <i>stall-in-progress?</i> .....	65
Possible implications for DOD.....	69
<b>Moving From “Understanding” to Operationalizing AI</b> .....	<b>71</b>
Theories of general <i>human</i> intelligence.....	71
Cattell–Horn–Carroll (CHC) Framework .....	74
Toward a common language.....	77
Observe-orient-decide-act (OODA).....	78
One possible bridge to help DOD pave a path from “understanding” to <i>operationalizing AI</i> .....	81
<b>Summary and Conclusions</b> .....	<b>89</b>

<b>Appendix A: “AI with AI” Podcast Mindmaps .....</b>	<b>95</b>
<b>Appendix B: Mindmap of JAIC Milestones 2017-2010 .....</b>	<b>111</b>
<b>Appendix C: Mindmap of DARPA AI-related Programs 2017-2010 .....</b>	<b>113</b>
<b>Appendix D: COVID-19 &amp; AI Mindmap.....</b>	<b>115</b>
<b>Appendix E: AI/ML Approaches, Methods, and Algorithms Taxonomy .....</b>	<b>117</b>
<b>Appendix F: AI Gaps and Limitations.....</b>	<b>119</b>
<b>Appendix G: Neuro-Lego World .....</b>	<b>121</b>
<b>Appendix H: Cattell-Horn-Carroll (CHC) Taxonomy of Cognitive Abilities .....</b>	<b>123</b>
<b>Appendix I: Mindmap of Possible Military Applications of AI .....</b>	<b>125</b>
<b>Figures .....</b>	<b>127</b>
<b>Abbreviations.....</b>	<b>129</b>
<b>References: Surveys of AI/ML Research .....</b>	<b>131</b>

# Introduction

---

In January 2017, CNA published a 300-page report, *AI, Robots, and Swarms* (hereafter referred to as the “2017/AI” paper),<sup>5</sup> that examines the conceptual, technical, and operational challenges facing the Department of Defense (DOD) as it pursues artificial intelligence (AI)-based technologies. The goal of this “quick look” follow-up paper is threefold:

1. Assess how the state-of-the-art (SOTA) in AI and machine learning (ML) has progressed since the publication of 2017/AI.
2. Revisit 2017/AI’s main findings—particularly those hinging on a set of core challenges associated with developing, adapting, and applying AI and ML to military problems—to see what remains relevant, what needs to be modified or discarded, and what new themes and issues have emerged between 2017 and 2020.
3. Recommend actions that DOD can take to help facilitate a transition from today’s era of “low-hanging fruit” applications of AI/ML (leveraged mostly to augment existing *human-centric* military capabilities) toward an era destined to usher in heretofore unimagined *human-centric*, *AI-centric*, and hybrid/symbiotic *human-AI* military systems and (even more broadly based) ecosystems.

The first two goals are achieved by surveying the most recent research literature published in science and technical journals, conferences, and preprint servers. More than 90 percent of the 150-plus research surveys that were used for this study (and that appear in the References section at the end of this paper) were published after 2017, and more than 70 percent were published no earlier than 2019. The proceedings from major AI-related conferences and symposia were also used—for example, AAAI Conference on Artificial Intelligence, ACM/IEEE International Conference on Human-Robot Interaction, and Agents and AI CogX, Conference on Computer Vision and Pattern Recognition (CVPR), among about a dozen others.<sup>6</sup>

---

<sup>5</sup> Andrew Ilachinski, *AI, Robots, and Swarms: Issues, Questions, and Recommended Studies*, CNA, DRM-20170U-01496, January 2017,

[https://www.cna.org/CNA\\_files/PDF/DRM-2017-U-014796-Final.pdf](https://www.cna.org/CNA_files/PDF/DRM-2017-U-014796-Final.pdf).

<sup>6</sup> There are many compilations of major AI and ML conference proceedings. One of the most complete repositories, ranked according to the “h5 index” (i.e., the largest number *h*, such that *h* papers published in the last five years have at least *h* citations each), is provided by Guide2Research, <http://www.guide2research.com/topconf/machine-learning>.

Additionally, we relied extensively on one of CNA's in-house resources—namely, the archive of CNA's weekly "AI with AI" podcasts.<sup>7</sup> Since its inaugural episode on November 3, 2017, a new podcast has been produced every week, with almost no breaks. As of September 2020, a total of 147 episodes have been produced, most of which are 30–40 minutes long (the longest is close to an hour). For the purposes of this study, the most significant part of this resource is the 2,600-plus pages of detailed notes that accompany the podcasts.<sup>8</sup> The podcast notes are a treasure trove of information insofar as they contain summaries of more than 400 AI/ML-related research studies (including most of the "breakthrough" and "milestone" events that have occurred since mid-2017) and more than 125 AI-related news stories that bear directly on AI/ML technologies (such as the recent spate of pushbacks against facial recognition software, discussed in later sections of this paper). The ready availability of this one in-house resource vastly simplified what otherwise would have been a tedious search for unifying themes and patterns in AI research. It also made the narrative summarized in these pages much easier to weave.

Offered as an information-rich resource to interested readers, a complete set of three-level-deep, time-ordered mindmaps of all research and news stories discussed on the "AI with AI" podcast (through August 2020), including contextually embedded hot-links to original source material, appears in **Appendix A: "AI with AI" Podcast Mindmaps**.

To accomplish the third goal, DOD needs to move beyond the stage of adjudicating principles, investment strategies, and policy options—the National Security Commission on Artificial Intelligence's (NSCAI) most recent report to Congress<sup>9</sup> offers more than 30 recommendations for adapting US national security policies and practices to rapidly evolving AI technologies—and of applying AI to military problems in piecemeal fashion to develop a consistent, coherent, holistic strategy for fully operationalizing AI's as-yet largely untapped (not of the "low-hanging fruit" variety) potential military applications. As a step toward this goal, we present a "template of a framework" that may help bridge the stubbornly persistent gap between *understanding* AI as a heavily technical research enterprise (under the purview of AI/ML researchers and developers)—replete with "the devil is in the details" level of basic challenges and limitations—and *operationalizing* its potential military applications (the wont of DOD portfolio managers, concept of operations developers, and senior decision-makers, who may lack the requisite technical training necessary to separate the wheat from the chaff).

A "bonus" section of the paper summarizes the response of the AI/ML research community to the ongoing (as of this writing, September 2020) COVID-19 pandemic.<sup>10</sup> While the inclusion of

---

<sup>7</sup> "AI with AI," CNA podcast, <https://www.cna.org/CAAI/audio-video>.

<sup>8</sup> The "AI with AI" notes are compiled by the author of this white paper, who also co-hosts the podcast itself.

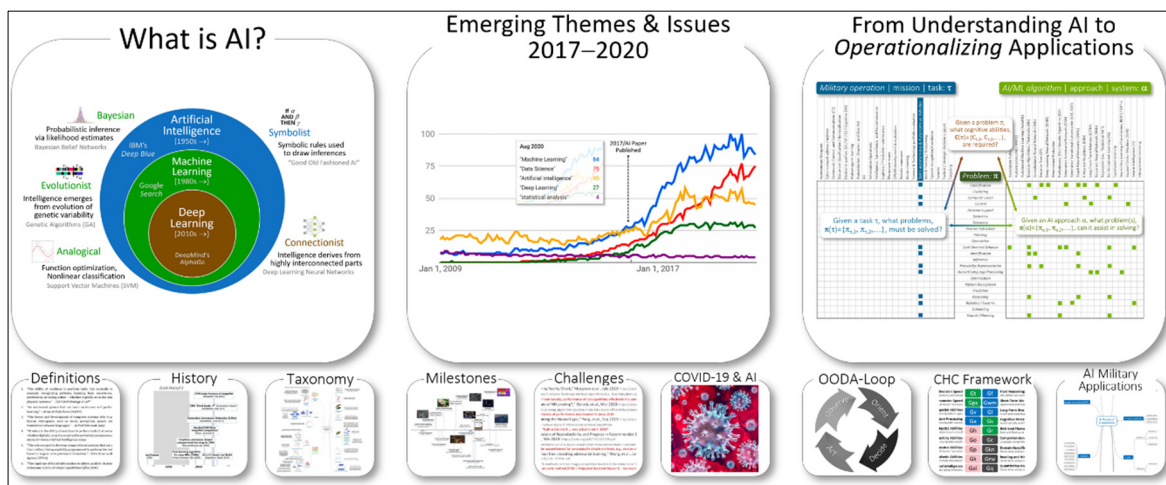
<sup>9</sup> National Security Commission on Artificial Intelligence, *Second Quarter Recommendations*, August 2020. [http://files2.dlapiper.com/DLA\\_Piper\\_Web\\_Images\\_US\\_2/pdfs/NSCAI%20Q2%20Memo\\_20200722.pdf](http://files2.dlapiper.com/DLA_Piper_Web_Images_US_2/pdfs/NSCAI%20Q2%20Memo_20200722.pdf).

<sup>10</sup> John Hopkins University of Medicine, "Coronavirus Resource Center," <https://coronavirus.jhu.edu/>.

this material may seem odd and off-topic, the welcome application of a technology seemingly tailor-made to deal with the complexities of an “emerging new threat” (albeit biological, not military) provides a veritable textbook of "lessons learned" for DOD as it looks to increasingly invest in and embrace new AI technologies.

## Organization of this paper

Figure 1. Visual schematic of the major sections of this paper



Source: CNA.

Figure 1 gives a visual overview of the three main parts of this white paper, which are detailed below:

1. **What is AI?** This section compares (and highlights the lack of consensus between) how the academic research community defines AI and how DOD defines it; provides a short history of AI, ML, and deep learning, highlighting two themes that are important for later parts of the paper—namely, several "dark periods" in AI research that occurred in prior decades and the perennially persistent challenges facing basic AI research; and concludes with two complementary views of AI—one as a categorical taxonomy of algorithms, the other as a field of scientific discovery.
2. **Emerging Themes and Issues.** This section of the paper summarizes the progress that has been made in AI and ML since the publication of “2017/AI” paper. Much of the material is culled from the in-house archive of notes summarizing over 400 research efforts discussed on CNA's "AI with AI" podcast (including “breakthrough” and “milestone” achievements, AI’s "misses," new research challenges, and the growing pushback against AI/ML). The section also includes a "bonus" discussion of how the AI/ML research community has responded to the COVID-19 pandemic (along with

"lessons learned" for DOD), and concludes with evidence that suggests that AI/ML may be entering (or has already entered) an era of "diminishing returns."

3. **From Understanding AI to Operationalizing Applications.** The final section introduces a "template of a framework" designed to help bridge the gap between *understanding* AI and *operationalizing* its military applications. This template leverages, and combines, three different frameworks: an AI method ↔ application matrix, the military-oriented OODA loop (described later), and a taxonomy of human cognitive abilities.

Two shorter (but no less substantive) sections not shown in Figure 1 are: (1) **Background**, which "sets the stage" with brief summaries of the US government's and DOD's recent AI investments, the establishment of the Joint Artificial Intelligence Center (JAIC), and several important AI ethics-related events and trends that have appeared since 2017; and (2) **Appendices A-I**, which provide a "self-contained" information resource that features 24 mindmaps organized around a variety of study-related topics (e.g., taxonomies of AI and ML methods and algorithms; recent breakthroughs and milestones; and gaps, challenges, and limitations of basic AI research). Each of these mindmaps assumes the form of a high-resolution Adobe PDF file (when opened in, say, Adobe Acrobat Reader or Adobe Acrobat Pro) that collectively contain an additional 800-plus embedded hot-link references.

# Background

---

## AI investments

The Department of Defense's (DOD) unclassified investments in AI have increased from about \$600 million in FY 2016 to \$927 million in FY 2020,<sup>11</sup> with more than 600 active AI-related projects (as of August 2019).<sup>12</sup> DOD has also requested \$800 million in FY 2021 to continue “the AI pathfinders, JAIC, and advanced image recognition (Project Maven)” and an additional \$1.7 billion for autonomy.<sup>13</sup>

A recent report on how AI/ML-related methods are used in nonmilitary and intelligence agencies (e.g., excluding DOD, the National Security Agency, and other agencies working on cyber-defense) reviews AI use at the 142 most significant federal departments, agencies, and sub-agencies.<sup>14</sup> Among the report's main findings are the following: (1) nearly half of the federal agencies studied (45 percent) have experimented with AI and related ML tools; (2) despite wide agency embrace of AI, the government still has a long way to go (only 12 percent of the techniques deployed were rated as “high” in sophistication); (3) AI poses deep accountability challenges (the report emphasizes AI's inherent lack of “explainability,” a theme we will also revisit); (4) while many agencies rely on private contractors to develop their AI applications, a majority of profiled use cases (53 percent) are products of in-house efforts by agency technologists; and (5) AI has the potential to raise distributive justice concerns and fuel political anxieties.

## Joint Artificial Intelligence Center (JAIC)

Arguably, the most significant AI-related event during 2017–2020 from a DOD perspective is the establishment of the JAIC.<sup>15</sup> The background for this event may be traced back to October 2016, when the (then) newly established Defense Innovation Board (DIB) released its first set

---

<sup>11</sup> Office of the Under Secretary of Defense (Comptroller)/Chief Financial Officer, Defense Budget Overview: United States DOD Fiscal Year 2020 Budget Request, March 2019, p. 1-9.

<sup>12</sup> Brendan McCord, “Eye on A.I.,” August 28, 2019, transcript available at <https://static1.squarespace.com/static/5b75ac0285ede1b470f58ae2/t/5d6aa8edb91b0c0001c7a05f/1567>

<sup>13</sup> Office of the Under Secretary of Defense (Comptroller)/Chief Financial Officer, “Defense Budget Overview: United States DOD Fiscal Year 2021 Budget Request,” February 2020 (Revised May 13, 2020), pp. 1–9.

<sup>14</sup> David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar, *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, February 2020. <https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf>.

<sup>15</sup> JAIC homepage: <https://www.ai.mil/>.

of recommendations.<sup>16</sup> Recommendation #5 (“Catalyze Innovations in Artificial Intelligence and Machine Learning”) was to establish a (at the time, unnamed) “centralized, focused, well-resourced organization” within DOD “to propel applied research in AI and machine learning.” Section 238 of the FY 2019 National Defense Authorization Act (NDAA), released in August 2018, called directly for the establishment of the Joint Artificial Intelligence Center (JAIC) to coordinate DOD projects of over \$15 million.<sup>17</sup>

JAIC was formally established in a June 2018 memo by Deputy Defense Secretary Patrick Shanahan<sup>18</sup> and charged with developing a common set of AI “standards ... tools, shared data, reusable technology, processes, and expertise” for the entire Defense Department. JAIC’s focus is on national mission initiatives (NMIs), which are broad, cross-functional programs that impact more than one mission or agency (e.g., predictive maintenance, humanitarian aid and disaster relief, cyberspace, and automation).<sup>19</sup> JAIC’s core mission themes (as articulated by JAIC’s first director, Lt. Gen. John N.T. “Jack” Shanahan) consist of<sup>20</sup> (1) accelerating delivery and adoption of AI capabilities across DOD; (2) establishing a common foundation for scaling AI’s impact (principally via by establishing an enterprise framework with a focus on shared data repositories for tools, standards, and cloud services); and (3) synchronizing all DOD AI and ML-related activities and projects.

In August 2019, JAIC launched a new marquee program for FY 2020 to use AI for “maneuver and fires,” a move that represents a move toward developing AI/ML tools for direct warfighting applications (e.g., operations-intelligence fusion, joint all-domain command and control, accelerated sensor-to-shooter timelines, autonomous and swarming systems, target development, and operations-center workflows).<sup>21</sup> In early September, JAIC announced that it is seeking an ethicist to help oversee military AI.<sup>22</sup> In October 2019, JAIC finally unveiled its public website.<sup>23</sup>

---

<sup>16</sup> The DIB is an advisory body to senior leadership in DOD and consists of representatives from the private sector, academia, and nonprofit organizations. The DIB’s 2017 set of recommendations may be accessed at [https://media.defense.gov/2017/Dec/18/2001857962/-1/-1/0/2017-2566-148525\\_RECOMMENDATION%2012\\_\(2017-09-19-01-45-51\).PDF](https://media.defense.gov/2017/Dec/18/2001857962/-1/-1/0/2017-2566-148525_RECOMMENDATION%2012_(2017-09-19-01-45-51).PDF). Other pre-2017 events appear in 2017/AI.

<sup>17</sup> John S. McCain National Defense Authorization Act for FY 2019, Public Law 115-232, August 13, 2018, <https://www.congress.gov/115/plaws/publ232/PLAW-115publ232.pdf>.

<sup>18</sup> Deputy Secretary of Defense Memorandum, *Establishment of the JAIC*, June 27, 2018, [https://admin.govexec.com/media/establishment\\_of\\_the\\_joint\\_artificial\\_intelligence\\_center\\_osd008412-18\\_r....pdf](https://admin.govexec.com/media/establishment_of_the_joint_artificial_intelligence_center_osd008412-18_r....pdf).

<sup>19</sup> Summary of the 2018 DOD AI Strategy: <https://fas.org/man/eprint/dod-ai.pdf>.

<sup>20</sup> Terry Moon Cronk, “DOD Unveils Its Artificial Intelligence Strategy,” Defense.Gov, Feb. 12, 2019.

<sup>21</sup> Justin Doubleday, “Pentagon plans to push AI into battle with new ‘maneuver and fires’ program,” Inside Defense, Aug. 30, 2019, <https://insidedefense.com/daily-news/pentagon-plans-push-ai-battle-new-maneuver-and-fires-program>.

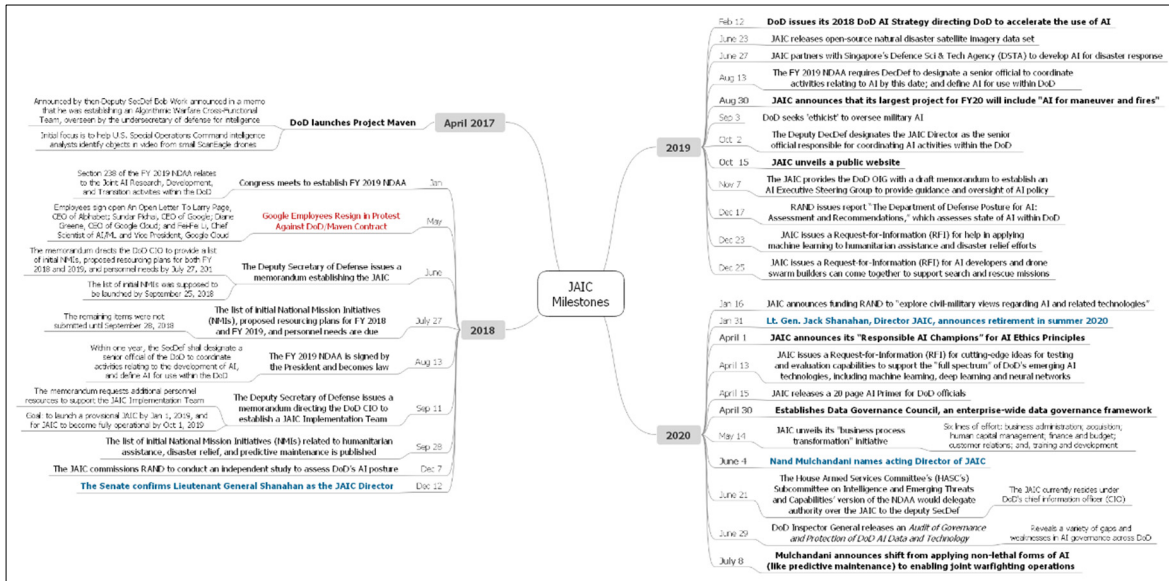
<sup>22</sup> C. Todd Lopez, “DOD Seeks Ethicist to Guide Artificial Intelligence Deployment,” Defense.Gov, Sept. 3, 2019, <https://www.defense.gov/Explore/Features/story/Article/1950724/dod-seeks-ethicist-to-guide-deployment-of-artificial-intelligence/source/GovDelivery/>.

<sup>23</sup> JAIC homepage, <https://www.ai.mil/>.



Figure 2 shows a time-ordered mindmap of significant JAIC-related milestones between 2017 and July 2020. **Appendix B** (of the PDF version of this document) contains a high-resolution version, and also includes embedded hot-link references to associated resources.

Figure 2. A mindmap/timelines of significant JAIC-related milestones, 2017–July 2020



Source: CNA.

JAIC’s most recent activities include (1) plans to ensure implementation of the DOD AI ethical principles (see the section on “AI ethics” below) by launching a cross-functional team of ethics advocates (announced in April 2020);<sup>24</sup> (2) establishing a data governance council with the goal of developing an enterprise-wide data governance framework;<sup>25</sup> and (3) naming Nand Mulchandani as acting director of JAIC on June 4, 2020 (following Lt. Gen. Shanahan’s announced retirement). In July 2020, Mulchandani announced a major forthcoming shift from applying nonlethal forms of AI (like predictive maintenance) to enabling joint warfighting operations.<sup>26</sup>

On June 29, 2020, the DOD Inspector General (DODIG) released an audit to determine DOD’s progress in developing an AI governance framework and standards. The audit revealed gaps and weaknesses in DOD’s enterprise-wide AI governance that are ostensibly JAIC’s responsibility and determined that JAIC had not yet developed a department-wide AI governance framework (having missed a March 2020 deadline for its release). Among the

<sup>24</sup> JAIC, April 1, 2020, [https://www.ai.mil/blog\\_04\\_01\\_20-shifting\\_from\\_principles\\_to\\_practice.html](https://www.ai.mil/blog_04_01_20-shifting_from_principles_to_practice.html).

<sup>25</sup> JAIC, April 30, 2020, [https://www.ai.mil/blog\\_04\\_30\\_20-jaic-leaders-establish-data-governance-council.html](https://www.ai.mil/blog_04_30_20-jaic-leaders-establish-data-governance-council.html).

<sup>26</sup> Nathan Strout, “Pentagon AI center shifts focus to joint war-fighting operations,” *Defense News*, July 8, 2020.

DODIG's recommendations for JAIC:<sup>27</sup> (1) determine a standard definition of AI and regularly (at least annually) consider updating the definition; (2) improve data sharing; and (3) develop a process to accurately track AI programs.

The JAIC also made one other announcement that is particularly significant for the purposes of this white paper: in April 2020, it released a 20-page AI primer for DOD officials.<sup>28</sup> Designed for DOD officials, the primer's self-expressed goal is to provide "simple answers to basic questions" like "What is AI?", "How does AI work?", and "Why is now an important time for AI?" What makes the primer's content significant for our purposes are the assertions that conclude the document's "Purpose" section:

Contrary to popular belief, you do not need to understand advanced mathematics or know computer programming languages to be able to answer the above questions accurately and to develop a practical understanding of AI relevant to your organization's needs. This guide will cover everything that the vast majority of DOD leaders need to know.<sup>29</sup>

For reasons that will become increasingly clear throughout this white paper, **we strongly disagree with these views**. AI is a far too complex endeavor—laden with myriad conceptual, mathematical, and programming *devil in the details*—level of technical issues and challenges—to be summarized with short descriptions of basic methods using charts and simple schematics. Moreover, the AI primer devotes less than a single page to discussing AI's limitations, technical challenges, and ethical issues and concerns. DOD officials who are already well versed (and technically trained) in the concepts covered in the primer obviously have no need for it. On the other hand, DOD officials who are *not* technically trained, or those who may be but are not yet knowledgeable about AI and/or ML—but are nonetheless responsible for making decisions involving military applications of AI and ML—stand little to gain of substantive value from JAIC's primer in our view. Still, this group of DOD officials obviously needs some resource that can provide the requisite technically relevant information on which they need to base policy decisions—presumably, one that does not involve taking a full graduate-level course on AI.

To be sure, there is an alarming *gap of mutual understanding* that exists between, on the one hand, technically trained practitioners of AI/ML (researchers, developers, and programmers)—who know the science but are not as well versed in the nuances of the applicability of AI/ML-derived methods to military problems—and, on the other hand, DOD

---

<sup>27</sup> Audit of Governance and Protection of Department of Defense Artificial Intelligence Data and Technology, DOD Office of Inspector General, DODIG-2020-098, July 1, 2020, <https://media.defense.gov/2020/Jul/01/2002347967/-1/-1/1/DODIG-2020-098.PDF>.

<sup>28</sup> Greg Allen, JAIC Chief of Strategy and Communications, *Understanding AI Technology*, April 2020, <https://www.ai.mil/docs/Understanding%20AI%20Technology.pdf>.

<sup>29</sup> *Ibid.*, p. 5.

officials who are obviously experts at adjudicating military requirements, but who may also lack the requisite depth of knowledge to make rigorously AI/ML-science-based policy and science and technology (S&T) portfolio decisions. **JAIC’s AI primer does not fill this gap;** instead, it merely highlights its existence (by the relative paucity and shallowness of the technically relevant information it provides). We will later introduce a “template of a framework” as a partial solution to closing this gap.

Referring back to Section 238 of the FY 2019 NDAA, DOD was also directed to publish a strategic roadmap for AI development and deployment, which was published in February 2019,<sup>30</sup> a day after a White House executive order announced the “American AI Initiative.”<sup>31</sup> The executive order directs the federal government to pursue five pillars for advancing AI: (1) invest in AI research and development (R&D), (2) unleash AI resources, (3) remove barriers to AI innovation, (4) train an AI-ready workforce, and (5) promote an international environment that is supportive of American AI innovation and its responsible use.

Also noteworthy is that although the FY 2019 NDAA directs the Secretary of Defense to produce a definition of AI by August 13, 2019, not only did this not happen, but—as of this writing (September 2020)—**no official US government definition of AI exists** (a fact, and the significance of which, we will discuss in the next section).

## Defense Advanced Research Projects Agency (DARPA)

Where JAIC is effectively tasked with delivering, adopting, and deploying AI, DARPA’s goal is to develop new AI technologies and make them operationally ready.

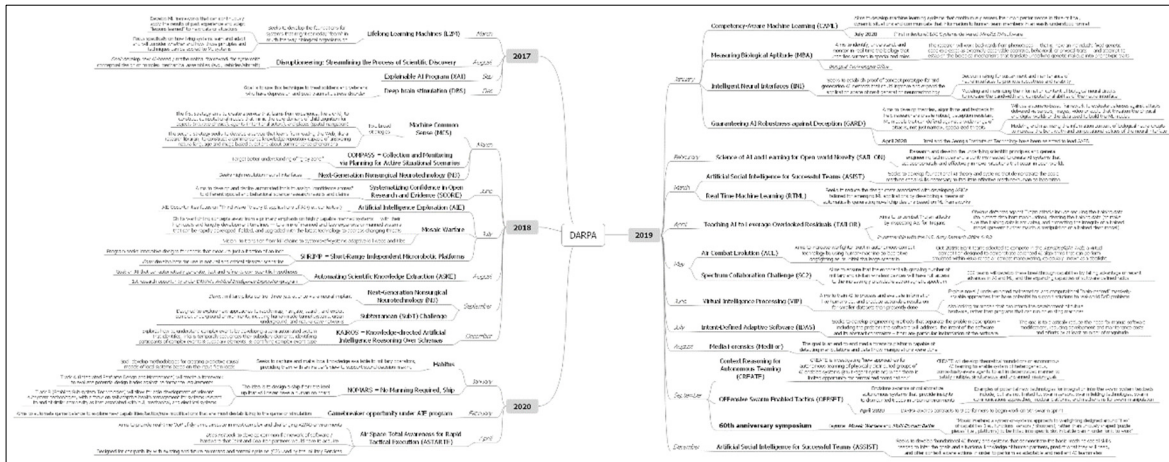
Figure 3 shows a time-ordered mindmap of significant DARPA AI-related milestones between 2017 and July 2020. **Appendix C** contains a high-resolution version of the mindmap shown in Figure 3, making it easier to see, and which also includes embedded hot-link references that automatically open up to accompanying resources.

---

<sup>30</sup> Summary of the 2018 DOD AI Strategy: <https://fas.org/man/eprint/dod-ai.pdf>.

<sup>31</sup> Executive Order 13859, “Maintaining American Leadership in Artificial Intelligence,” Executive Office of the President, Feb. 11, 2019, <https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence>.

Figure 3. A mindmap/timelines of DARPA AI-related program announcements, 2017–July 2020



Source: CNA.

## AI ethics

Although the focus of this paper is not on ethics per se, we would be remiss not to at least highlight a few significant events and trends that have appeared between 2017 and 2020. “Ethics” plays an important role later in this paper as a backdrop to a discussion on the burgeoning pushback against the development of certain AI technologies.

In the last few years, numerous studies have been published on AI ethics—including concept papers, statements of principle, and developmental frameworks—from academics, private and public organizations, industry, and government institutions. There have been more than 30 ethics-related stories covered on the “AI with AI” podcast in that time.<sup>32</sup> Recent headlines include (1) civil liberties issues (namely, *privacy*) surrounding otherwise laudable crowdsourced attempts (by Apple, Google, and others) to keep tabs on the spread of COVID-19;<sup>33</sup> (2) a sweeping Justice in Policing Act that includes several provisions that hold police accountable while using body cameras and limiting the use of facial recognition technologies;<sup>34</sup> and (3) the first known case of someone being mistakenly arrested in the US due to facial recognition technology.<sup>35</sup>

<sup>32</sup> Search on “ethics” in “AI with AI” Podcast mindmaps in **Appendix A**.

<sup>33</sup> Ashkan Soltani, Ryan Calo, and Carl Bergstrom, “Contact-tracing apps are not a solution to the COVID-19 crisis,” TechStream, Brookings Institution, 27 April 2020, <https://www.brookings.edu/techstream/inaccurate-and-insecure-why-contact-tracing-apps-could-be-a-disaster/>.

<sup>34</sup> US House of Representatives, Committee on the Judiciary, *George Floyd Justice in Policing Act of 2020*, 116th Cong., 2nd Sess., H.R. 2120, <https://www.congress.gov/bills/116/congress-house-bill/7120/text>.

<sup>35</sup> Kashmir Hill, “Wrongfully Accused by an Algorithm,” *New York Times*, 24 June 2020.

Some notable AI ethics frameworks that have recently appeared include the following:

- **Executive Order on Maintaining American Leadership in AI**, which provides several broad ethical provisions for moving forward with developing AI technologies (but which, interestingly, does not actually include the word “ethics”).<sup>36</sup>
- **The Institute of Electrical and Electronics Engineers’ (IEEE) *Ethically Aligned Design, Version 2***, a 260-page document from IEEE’s Global Initiative on Ethics of Autonomous and Intelligent Systems that contains detailed discussions of engineering-level ethical considerations and offers more than 100 recommendations for technologists, policy-makers, and academics.<sup>37</sup>
- **European Commission High-Level Expert Group’s “Ethics Guidelines,”** the aim of which is to offer guidance on fostering and securing ethical and robust AI to promote AI’s trustworthiness. Notably, this document seeks to go beyond a list of ethical principles by providing specific guidance on how such principles can be *operationalized* in socio-technical systems.<sup>38</sup>
- **Organization for Economic Cooperation and Development’s (OECD) “AI Principles,”** which identifies five complementary values-based principles for the responsible stewardship of trustworthy AI (which were subsequently adopted by the Group of 20 in June 2019).<sup>39</sup>
- **The Defense Innovation Board’s (DIB) “AI Principles: Recommendations on the Ethical Use of AI by the Department of Defense,”** which encompasses five main areas ([AI must be...] *responsible, equitable, traceable, reliable, and governable*). These principles were adopted by DOD in February 2020.<sup>40</sup>

Keep in mind that the list above contains just the highlights. Indeed, so many “frameworks” and “statements of principle(s)” have appeared in recent years that “frameworks of frameworks” are becoming as common as the frameworks themselves. For example, a recent Harvard University study analyzed common threads among 36 separate ethical guidelines (organized around nine key themes: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional

---

<sup>36</sup> Executive Order 13859.

<sup>37</sup> IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2*, 2017, [http://standards.ieee.org/develop/indconn/ec/autonomous\\_systems.html](http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html).

<sup>38</sup> European Commission, High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, April 2019, <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>.

<sup>39</sup> Organization for Economic Cooperation and Development, *Going Digital, “OECD Principles on AI,”* <http://www.oecd.org/going-digital/ai/principles/>.

<sup>40</sup> Department of Defense, “DOD Adopts Ethical Principles for Artificial Intelligence,” Press Release, 24 Feb. 2020.

responsibility, promotion of human values, and international human rights);<sup>41</sup> the *AI Ethics Guidelines Global Inventory* maintains an on-line archive with more than 160 international AI ethics guidelines.<sup>42</sup>

To date, almost all frameworks are assertions of principles and only rarely give practical ideas on how to put these principles into practice. Among the exceptions are: (1) a report published by *The AI Ethics Impact Group* (AIEI)—an interdisciplinary European consortium—that “offers concrete guidance to decision-makers in organizations developing and using AI on how to incorporate values into algorithmic decision-making, and how to measure the fulfilment of values using criteria, observables, and indicators combined with a context-dependent risk assessment”,<sup>43</sup> and (2) the US intelligence community’s *AI Ethics Framework*, published in June 2020, which—though it is short, at only six pages—provides specific guidance on how to apply ethical considerations (achieved by listing specific questions that decision-makers must ask regarding basic issues such as goals and risks, legal obligations, judgment and accountability, and so on).<sup>44</sup>

Two final, recently published, internationally oriented “AI Ethics” surveys are (1) a study on issues and initiatives published by the European Parliamentary Research Service’s Panel for the Future of Science and Technology in March 2020,<sup>45</sup> and (2) a survey published by the Montreal AI Ethics Institute in June 2020.<sup>46</sup>

Much of the commentary above on AI ethics foreshadows the broader issue that lies at the heart of this occasional paper—namely, how to best *operationalize* AI methods and technologies in general. Viable frameworks for either issue<sup>47</sup> must all start by addressing the unique perspectives of (partly overlapping, but mostly distinct) stakeholders: *developers* (i.e., people getting their hands dirty with concepts, math, and computer code); *policy-makers* (those who transform abstract “principles” into actionable plans and/or—in the military domain—those

---

<sup>41</sup> Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center for Internet and Society, Harvard University, January 2020, [https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final\\_v3.pdf?sequence=1&isAllowed=y](https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf?sequence=1&isAllowed=y).

<sup>42</sup> “AI Ethics Guidelines Global Inventory,” AlgorithmWatch, April 2020, <https://inventory.algorithmwatch.org/>.

<sup>43</sup> AI Ethics Impact Group, *From Principles to Practice: An Interdisciplinary Framework to Operationalize AI Ethics*, April 2020, [https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO\\_2020\\_final.pdf](https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf).

<sup>44</sup> “Artificial Intelligence Ethics Framework for the Intelligence Community, Version 1.0,” June 2020, <https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community>.

<sup>45</sup> European Parliament, Panel for the Future of Science and Technology, *The Ethics of Artificial Intelligence: Issues and Initiatives*, European Parliamentary Research Service Report PE 634.452, March 2020, [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS\\_STU\(2020\)634452\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf).

<sup>46</sup> Abhishek Gupta, Camylle Lanteigne, Victoria Heath, Marianna Ganapini, Erick Galinkin, et al., *The State of AI Ethics*, Montreal AI Ethics Institute, June 2020, <https://arxiv.org/ftp/arxiv/papers/2006/2006.14662.pdf>.

<sup>47</sup> While “AI ethics” is obviously a subset of more general AI “issues,” for the purposes of this section’s discussion we treat these as separate “problems.”

responsible for adjudicating science and technology, or S&T, portfolios); and *end users* (people who actually use newly deployed capabilities and technologies). Complicating the issue (though more on the “ethics” side) is that, on the international stage, members of all three groups must adhere to local sets of social and cultural norms.

A recent effort to help bridge the divide between developers and end users is a survey by Microsoft that culminates in a machine-learning-practitioner-centric AI ethics checklist.<sup>48</sup> On the military stage, the JAIC in May 2020 announced its “Responsible AI Champions” for AI Ethics Principles, demonstrating an ostensible shift from principles to practice.<sup>49</sup> But efforts like these are nascent, and their effectiveness remains to be seen.

---

<sup>48</sup> Michael A. Madaio, Luke Start, Jennifer Wortman Vaughan, and Hanna Wallach, “Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI,” paper presented at the CHI Conference on Human Factors in Computing Systems, April 25–30, 2020, Honolulu, HI.

<sup>49</sup> “JAIC Completes Responsible AI Champions Pilot,” The JAIC, 8 July 2020, [https://www.ai.mil/blog\\_07\\_08\\_20-jaic\\_completes\\_responsible\\_ai\\_champions\\_pilot.html](https://www.ai.mil/blog_07_08_20-jaic_completes_responsible_ai_champions_pilot.html).

[This page intentionally left blank]



# What Is AI?

---

The phrase “artificial intelligence” (if not its meaning) was coined by John McCarthy at the 1956 British Dartmouth Summer Conference.<sup>50</sup> However, the conceptual rudiments of AI go back at least another decade to when the first artificial neuron (or “threshold logic unit”) was introduced in 1943 by McCulloch and Pitts.<sup>51</sup> Since then, the field has advanced—in fits and starts (a few milestones of which are highlighted later in this section)—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. It is the latter approach, exemplified by a broad class of “deep learning” methods—particularly those able to learn from vast datasets without (or with only minimal) human supervision—that has garnered widespread attention in recent years.<sup>52</sup>

## Definitions

Despite its long history and increasingly rapid advances, there has never been a universally agreed-upon definition of what “AI” is—not in the academic research community, and not in DOD. Indeed, this persistent lack of agreement—a lack of even a *consistency* in approaches to defining AI—only exacerbates the aforementioned divide between those who develop AI and those (in DOD) who are charged with applying and operationalizing its potential applications. For the sake of brevity, and because the singular goal of this occasional paper is to present a roadmap for what we earlier called a “template of a framework” toward bridging this divide, the remaining parts of this section introduce only those aspects of AI and ML—including a unique perspective not found in other expositions, so far as this author is aware—that further the main narrative. Many more details (and references) may be found in the 2017/AI paper as well as in the mindmaps in **Appendices A–I**.

---

<sup>50</sup> John McCarthy is generally acknowledged as one of the “founding fathers” of artificial intelligence, along with Marvin Minsky (with whom he worked at MIT), Allen Newell, and Herbert Simon; Sam Williams, *Arguing A.I.: The Battle for Twenty-first-Century Science*, New York: AtRandom.com Books, 2002.

<sup>51</sup> W. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bulletin of Mathematical Biophysics* 5 (1943).

<sup>52</sup> Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2016, <https://www.deeplearningbook.org/>.

## AI research community definitions of AI: *little consensus*

On the academic and research side, Norvig and Russell, in the introduction to their opus *Artificial Intelligence: A Modern Approach*,<sup>53</sup> provide a multitude of disparate definitions of AI from other standard textbooks, some of which stress “thought processes,” others “reasoning,” and still others “rationality.”<sup>54</sup> For example, AI may be defined as:

- “Automation of activities associated with human thinking, such as decision-making, problem solving, learning”
- “Study of mental faculties through the use of computational models”
- “Study of the computations that make it possible to perceive, reason, and act”
- “The art of creating machines that perform functions that require intelligence when performed by people”
- “Study of how to make computers do things that, at the moment, people are better at”
- “A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes”
- “Branch of computer science concerned with the automation of intelligent behavior.”

Nilsson, in the preface of his 550-page history of the field—*Quest for Artificial Intelligence*<sup>55</sup>—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:<sup>56</sup> “AI is the field that studies the synthesis and analysis of computational agents that act intelligently.”

Norvig and Russell organize these (and other) definitions according to the degree to which they tend to emphasize *cognition* over *action*—that is, AI are systems that (1) “**act like humans,**” which includes the well-known “Turing Test” approach to determine the veracity of an AI system, wherein a human decides whether the “conversation” he is asked to engage in is with another human or an AI<sup>57</sup> (this category also includes natural language processing, knowledge representation, automated reasoning, and robotics); (2) “**think like humans,**”

---

<sup>53</sup> Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., New York: Pearson, 2020.

<sup>54</sup> The authors point out that proponents of these various approaches have both helped and disparaged one another.

<sup>55</sup> Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, New York: Cambridge University Press, 2009.

<sup>56</sup> David L. Poole and Alan K. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*, New York: Cambridge University Press, 2010.

<sup>57</sup> A. M. Turing, “Computing Machinery and Intelligence,” *Mind* 49 (October 1950): 433–460.

which includes cognitive modeling, introspection, psychological experiments, and brain imaging; (3) “**act rationally**,” which is essentially a symbolic processing–driven “laws of thought” approach and includes probabilistic reasoning and logic-based methods; and (4) “**think rationally**,” which is effectively a “rational agent” approach that entails developing methods for “best outcome” decision-making under uncertainty and limited rationality.

The Holy Grail of AI is artificial *general* intelligence (AGI)<sup>58</sup>—that is, an AI that demonstrates human-level intelligence across the same broad range of cognitive tasks in which humans are proficient. The consensus among the AI/ML research community is that (as of 2020), there is no clear path to achieving AGI using today’s algorithms.<sup>59</sup> A survey conducted at the 2012 Singularity Summit of AI experts (an annual conference of the Machine Intelligence Research Institute, founded in 2006 at Stanford University) found a wide range of predicted dates for when AI will be equal to or surpass human-level general intelligence, with a median date range in the year 2040.<sup>60</sup> Instead, today’s AI is best described as *narrow* artificial intelligence insofar as the cognitive tasks and problems for which it is best suited, even as it seemingly routinely outperforms humans on specific tasks, is limited in scope. An otherwise superhuman-performing algorithm such as AlphaGo (that defeated world champion Go player Lee Sedol in 2016<sup>61</sup>) could not only not play a different game (say, chess), but would also have to be retrained to play Go if the game’s board had been minimally changed to include, say, a single additional row or column.

## DOD definitions of AI: *even less consensus*

Getting back to the many definitions but little consensus about “what AI is” among researchers, there is even greater disparity within DOD and other US government agencies. Recall (from the previous section) that although the 2019 NDAA mandated the Secretary of Defense to produce a definition of AI by 13 August 2019, this did not happen. Indeed, as of this writing (September 2020), *no official US government definition of AI exists*. Notably, the 2019 NDAA *itself* provides a “working definition” of AI, asserting that “AI includes...

---

<sup>58</sup> AGI was introduced as the *title* to a conference held in 2006; *Artificial General Intelligence*, ed. Ben Goertzel and Cassio Pennachin, New York: Springer-Verlag, 2006.

<sup>59</sup> Gary Marcus and Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*, New York: Pantheon Books, 2019; Melanie Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*, London: Pelican Books, 2019.

<sup>60</sup> Stuart Armstrong and Kaj Sotola, “How We’re Predicting AI—Or Failing To,” in *Beyond AI: Artificial Dreams*, ed. Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, Pilsen, Czech Republic: University of West Bohemia Research Center, 2012.

<sup>61</sup> Christoph Koch, “How the Computer Beat the Go Master,” *Scientific American*, 19 March 2016, <https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/>.

1. Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to datasets
2. An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action
3. An artificial system designed to think or act like a human, including cognitive architectures and neural networks
4. A set of techniques, including machine learning, that is designed to approximate a cognitive task
5. An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.”

Although this definition is laudable as a draft—it covers a lot of ground, but was intended by the 2019 NDAA to be used only as a stepping stone to more refined future versions—its sweeping nature makes it hard to use as a foundation on which to base investment strategies or policy positions. The DIB gave the 2019 NDAA’s working definition a scorching critique, stating that “each definition individually carries with it policy and ethical challenges, and attempting to incorporate them all yields an unwieldy definition of AI.”<sup>62</sup> The DIB argues that if AI is defined either too narrowly or too broadly, the scope of AI capabilities may either be overly constrained or fail to specify the unique capacity that AI methods may offer. With this view in mind, DIB’s specific criticisms include definition #1 on the 2019 NDAA list (which DIB asserts unnecessarily conflates autonomous systems and AI systems, “leaving open questions [about] systems that are not dynamic in their design and architectures or non-learning systems”<sup>63</sup>); definition #4 (which, by limiting AI to cognitive tasks, “may also limit various uses of AI for physical task completion, where the repetitive physical task requires very little cognitive abilities”<sup>64</sup>) and definitions #2–5 combined (which, by involving “human-like” capacities, may be unnecessarily limiting).

The DIB itself offers this more succinct and open-ended definition of AI:

[The DIB] considers AI to be a variety of information processing techniques and technologies used to perform a goal-oriented task and the means to reason in

---

<sup>62</sup> Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, Supporting Document, October 2019, [https://admin.govexec.com/media/dib\\_ai\\_principles\\_-\\_supporting\\_document\\_-\\_embargoed\\_copy\\_\(oct\\_2019\).pdf](https://admin.govexec.com/media/dib_ai_principles_-_supporting_document_-_embargoed_copy_(oct_2019).pdf).

<sup>63</sup> *Ibid.*, p. 9.

<sup>64</sup> *Ibid.*, p. 8.

the pursuit of that task. These techniques can include, but are not limited to, symbolic logic, expert systems, machine learning (ML), and hybrid systems.<sup>65</sup>

The DIB justifies this formulation on the grounds that it is both consistent with how DOD has viewed, developed, and deployed AI systems over the past 40 years, and that it is flexible enough to allow finer-grained distinctions to be made, as needed, between legacy systems run on expert or hybrid systems and newer systems using ML methods. Importantly, the DIB also emphasizes that **AI is not synonymous with autonomy**, a point also emphasized in the 2017/AI paper. While some autonomous systems may depend (even critically) on AI/ML, this need not be generally the case. For example, while the 2012 DOD Directive (DODD) 3000.09 defines an autonomous weapon system (AWS) as one that “once activated, can select and engage targets without further intervention by a human operator,”<sup>66</sup> it does not specify *the manner in which* an AWS decides its actions. An AI method *may* be involved (DODD 3000.09 does not explicitly prohibit it), but it is not required.

To the 2019 NDAA’s and DIB’s already discordant definitions of AI, we can add the following definitions (promulgated by DOD and used by a few selected DOD agencies):<sup>67</sup>

1. “The ability of machines to perform tasks that normally require human intelligence – for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action – whether digitally or as the smart software behind autonomous physical systems”—*2018 DOD Strategy on AI*<sup>68</sup>
2. “An automated system that can learn on its own and perform multiple tasks using machine learning”—*Army AI Task Force (AAITF)*
3. “The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages”—*AI Portfolio lead, US Navy*
4. “AI refers to the ability of machines to perform tasks that normally require human intelligence... whether digitally or as the smart software behind autonomous physical systems”—*2019 US Air Force Artificial Intelligence Annex*
5. “The science/goal to develop computational systems that can reason about problems and solve them without being explicitly programmed to perform the task and that adapt to new situation based on exposure to previous information”—*Data Science AI Office, Defense Threat Reduction Agency (DTRA)*

---

<sup>65</sup> Ibid., p. 9.

<sup>66</sup> DOD Directive 3000.09, *Autonomy in Weapon Systems*, November 21, 2012 (incorporating Change 1, May 8, 2017), <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf>.

<sup>67</sup> Apart from the first definition on this list that is quoted from the 2018 DOD Strategy on AI, all other definitions are taken from Table 1 (p. 9) in *Audit of Governance and Protection of Department of Defense Artificial Intelligence Data and Technology*, DOD Office of Inspector General, DODIG-2020-098, 1 July 2020. <https://media.defense.gov/2020/Jul/01/2002347967/-1/-1/1/DODIG-2020-098.PDF>.

<sup>68</sup> Department of Defense, *Summary of the 2018 DOD AI Strategy: Harnessing AI to Advance Our Security and Prosperity*, February 2019, <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>.

6. “The rapid use of broad information to inform analytic decision making”—*Portfolio Leader for Autonomy and AI, Strategic Capabilities Office (SCO)*.

To be sure, this is a cacophony of ideas, principles, and varying emphases. DOD’s 2018 AI Strategy conflates *narrow* AI with AGI and subsumes AI in autonomy. The AAITF’s definition partly aligns AI with automation, conflates AI with ML (when ML is properly a *subset* of AI), and unnecessarily requires (presumably narrow AI, and not AGI) methods to be able to solve “multiple tasks.” And the SCO’s definition seems to conflate AI with basic data science (along with an unnecessarily overt constraint on the types of problems “AI” is used to solve); in this sense, it falls squarely on the other side of the 2019 NDAA’s “overly broad” definition.

## So what is AI, really?

We first approach the answer to the question above from the perspective of what AI is *not*. Indeed, our repeated use throughout this paper of the phrase, “*the devil is in the details*,” is a deliberate heuristic reminder to decision-makers and other DOD stakeholders (including warfighters) that, even apart from their mutual discordance, none of the definitions of AI (such as those appearing in the previous section) come close to encapsulating what it means, pragmatically, to develop, assess the applicability of, and ultimately deploy AI systems. To be even more blunt, the meaning of “AI” is not to be found in any “definition,” however pithy or eloquently stated.

This is because “AI” does not really refer to any *one* thing (or method, or class of “solutions” to a problem). Instead, it is best *described*—not defined—as a **continually evolving self-organized community of research activity focused on developing “methods of finding solutions to problems” and whose ultimate goal is to replace human cognition with some combination of hardware and software**. The “AI” of the 1960s is emphatically not the same as what “AI” was in the 1990s, and both are very different from what constitutes their past-decade and present-day cousins; yet the same generic term “AI” is used interchangeably to label the ostensibly same “AI things” spanning AI’s entire history. We must also be careful to disentangle AI as (1) a *field of scientific discovery*, wherein genuine breakthroughs and milestones (that are hard enough to recognize at the time of their discovery) compete with “fad of the day” ideas eventually forgotten; and (2) the top-most level of a *categorical taxonomy* of specific approaches, methods, and algorithms, which is how “AI” is commonly viewed (e.g., “AI is neural networks,” “AI is machine learning,” or “AI is reinforcement learning”).

As is true for science in general, discovering new concepts and methods for AI is hard and messy, and it is all but impossible to predict when the next “breakthrough” might occur or if past breakthroughs (e.g., a specific method designed to “solve” a given problem) can be adopted and/or scaled for more physically realistic and militarily relevant scenarios. Although we are not advocating the position that (nontechnically trained) stakeholders must take time

to earn advanced degrees in AI or ML, we are suggesting that no meaningful discourse can ensue on these subjects (including the adjudication of policy and designing S&T investment portfolios) unless and until a requisite level of methodological rigor is established.

The last part of this paper (see the **“Understanding vs Operationalizing AI”** section) introduces a conceptual framework for developing a *mutually translatable language* between AI/ML developers and practitioners, on one side, and military stakeholders and decision-makers, on the other. But before getting to this framework, we must first “set the stage.” The remaining parts of this section present two complementary “histories” of AI—one a conventional history of major milestones, the other a unique take on *how* some of these milestones happened—and summarize (as yet unsolved) technical challenges. The following section brings the historical narrative up to date by summarizing the state of progress from 2017 to 2020.

## Short history

Recall that progress in AI has advanced principally along two concurrent methodological perspectives: a *top-down* approach, in which knowledge about a specific problem domain is curated by human SMEs, codified by rules, and implemented in software (the goal being to reproduce “human-like” reasoning); and a *bottom-up* approach, which mimics nature’s evolutionary propensity to build “complex” structures out of “simpler” parts (exemplified by modern “deep learning” methods).<sup>69</sup>

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 an 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.<sup>70</sup>

Research in NLP goes back to the roots of computer science in the 1940s and 1950s.<sup>71</sup> 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

---

<sup>69</sup> This section is a shortened, but updated (and slightly altered) version of the material that appears on pp. 43–68 in Ilachinski, *AI, Robots, and Swarms*.

<sup>70</sup> Robin Deits, Stefanie Tellex, Pratiksha Thaker, Dimitar Simeonov, Thomas Kollar, et al., “Clarifying Commands with Information-Theoretic Human-Robot Dialog,” *Journal of Human-Robot Interaction* 2. no. 2 (2013): 58–79, doi: 10.5898/JHRI.2.2.Deits.

<sup>71</sup> Stephen Lucci and Danny Kopec, *Artificial Intelligence in the 21st Century*, Dulles, VA: Mercury Learning and Information, 2013.

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 a 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, and searching information resources that are relevant to a specific query or subject from a given source of documents); (3) *information extraction* (the extraction of semantic information from text); and (4) *deep learning* (discussed below).

## Machine learning (ML)

“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 (see the mindmind in **Appendix E** for a more complete taxonomy):<sup>72</sup>

- *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* (sometimes also called *hierarchical learning*), which refers to a class of ML algorithms designed to find multiple high levels of abstract representations of patterns in data; this is discussed in more detail in the following section);
- *Genetic algorithms* and other *evolutionary programming* techniques that mimic the dynamics of natural selection;<sup>73</sup>

---

<sup>72</sup> Russell and Norvig, *Artificial Intelligence: A Modern Approach*.

<sup>73</sup> Zbigniew Michalewicz and David B. Fogel, *How to Solve It: Modern Heuristics*, New York: Springer-Verlag, 2005.



- *Inductive logic programming*, designed to infer a hypothesis from a knowledge base and a set of positive and negative examples;<sup>74</sup>
- *Neural networks*, which are inspired by the structure and function of biological neural networks;<sup>75</sup>
- *Reinforcement learning*, which is inspired by behavioral psychology and refers to a technique whereby learning proceeds by adaptively constructing a sequence of actions that collectively maximize some long-term reward;<sup>76</sup> and
- *Support vector machines (SVM)*,<sup>77</sup> which are essentially multidimensional binary classification algorithms.

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—that is, 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 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.<sup>78</sup> 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

---

<sup>74</sup> Stephen H. Muggleton and Hiroaki Watanabe, eds., *Latest Advances in Inductive Logic Programming*, London: Imperial College Press, 2014.

<sup>75</sup> Mohamad Hassoun, *Fundamentals of Artificial Neural Networks*, Cambridge, MA: MIT Press, 2003.

<sup>76</sup> Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 1998.

<sup>77</sup> Nello Cristianini, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, New York: Cambridge University Press, 2000.

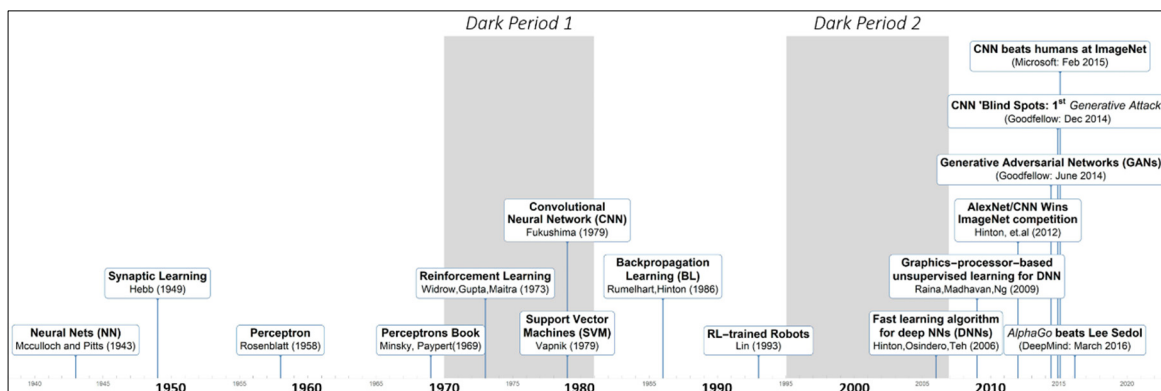
<sup>78</sup> Jurgen Schmidhuber, *Deep Learning in Neural Networks: An Overview*, Technical Report IDSIA-03-14, arXiv:1404.7828 v4 (2014).

renaissance in a slightly modified form known as *deep learning* (DL),<sup>79</sup> due mainly to the confluence of three factors: (1) exponential growth in computing power, (2) the exponentially dwindling cost of digital storage coupled with an exponential growth of available data, and (3) a new generation of fast learning algorithms for multilayer 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 4 shows a timeline of major milestones through 2017, starting with the first mathematical model of a neuron introduced in 1943 (the next section, “**Emerging AI Themes and Issues**,” brings this timeline up to the present time).<sup>80</sup>

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,  $x$ .

Figure 4. Timeline of milestones in the development of neural networks and deep learning



Source: CNA.

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).

<sup>79</sup> Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, Cambridge, MA: MIT Press, 2016.

<sup>80</sup> W.S. McCulloch and W.A. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bulletin of Mathematical Biophysics* 5 (1943): 115–133.

LR is essentially the idea behind the perceptron, introduced by Rosenblatt in 1958 as a mathematical distillation of how biological neurons operate.<sup>81</sup> 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 perceptrons 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 4), during which the funding for further research and the number of published papers dropped significantly from prior years, followed the landmark publication of the book *Perceptrons* in 1969.<sup>82</sup> The book argued, correctly, that because the Boolean exclusive-OR (or XOR) function is not linearly separable,<sup>83</sup> 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—that is, an NN in which there are *hidden layers* sandwiched between the input and output layers (see left-hand side of Figure 5a). But the only known learning algorithm at the time of the book’s publication applied only to the simplest single-input-layer/single-output-layer NNs. It was not until 1986 that the first “dark period” of NN research finally ended when the so-called *backpropagation learning* (BL) method, which could be applied to NNs with hidden layers, was introduced.<sup>84</sup>

---

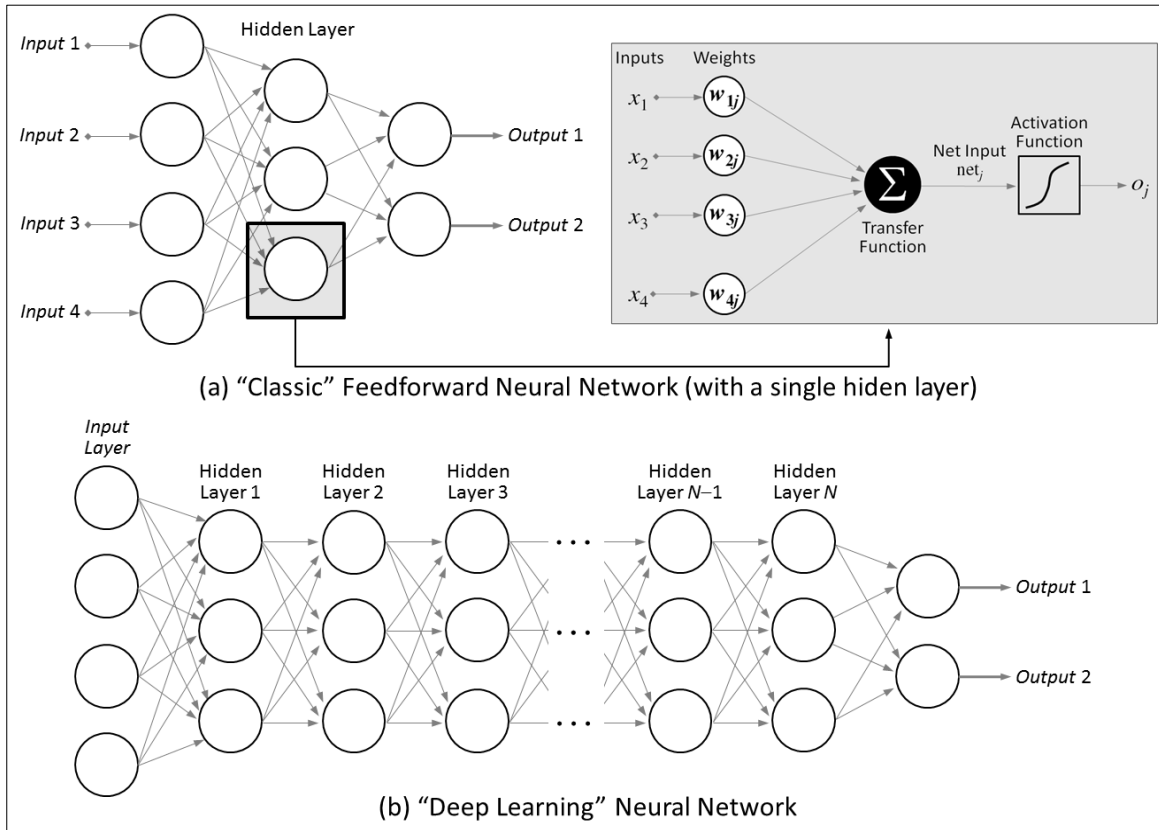
<sup>81</sup> F. Rosenblatt, “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain,” *Psychological Review* 65, no. 6 (1958).

<sup>82</sup> Marvin Minsky and Seymour A. Papert, *Perceptrons: An Introduction to Computational Geometry*, Cambridge, MA: MIT Press, 1969.

<sup>83</sup> 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. See Chapter 4 in Simon Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed., Upper Saddle River, NJ: Prentice Hall, 1999.

<sup>84</sup> David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, “Learning representations by back-propagating errors,” *Nature* 323 (1986): 533–536.

Figure 5. Schematic illustrations of neural network designs



Source: CNA.

Three years later, Hornik et al. proved that multi-layered NNs can learn any function, including XOR.<sup>85</sup> Also in 1989 was the first landmark application of BL to the automatic recognition of handwritten ZIP code numbers,<sup>86</sup> 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 3-D images, the layers of a CNN might be arranged in three dimensions (width, height, and depth).<sup>87</sup> Modern incarnations of CNNs include *pooling layers* (positioned in between convolutional layers), which effectively reduce the spatial

<sup>85</sup> Kurt Hornik, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural Networks* 2, no. 5 (1989): 359-366.

<sup>86</sup> Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation* 1, no. 4 (1989): 541-551.

<sup>87</sup> L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

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.

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.<sup>88</sup> Also around this time, an RL-based AI system—called TD-Gammon—famously “taught itself” to play backgammon at a superhuman level.<sup>89</sup> 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 4), which ended around 2006. The reason was that when TD-Gammon’s learning algorithm was applied to other (albeit more “complex”) games such as chess<sup>90</sup> and Go,<sup>91</sup> 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 circa 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 an NN that had many hidden layers<sup>92</sup>—and it is the presence of many hidden layers that is the cornerstone of most modern deep learning systems (see Figure 5b).<sup>93</sup>

Even as the raw power of available computers was steadily increasing until very recently (thanks to Moore’s Law<sup>94</sup>), it was not until both a “fast learning” algorithm for deep learning

---

<sup>88</sup> L. Lin, “Reinforcement Learning for Robots Using Neural Networks,” PhD thesis, Carnegie-Mellon University, School of Computer Science, CMU-CS-93-103, 1993.

<sup>89</sup> Gerald Tesauro, “Temporal difference learning and TD-Gammon,” *Communications of the ACM* 38, no. 3 (March 1995).

<sup>90</sup> Sebastian Thrun, “Learning to Play the Game of Chess,” *Advances in Neural Information Processing Systems* 7 (1995): 1069–1076.

<sup>91</sup> Nicol N. Schraudolph, Peter Dayan, and Terrence J. Sejnowski, “Temporal difference learning of position evaluation in the game of Go,” *Advances in Neural Information Processing Systems* 6 (1994).

<sup>92</sup> Chapter 10 in Andrew Ilachinski, *Cellular Automata: A Discrete Universe*, Hackensack, NJ: World Scientific Press, 2001.

<sup>93</sup> 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  $N^{\text{th}}$  hidden layer to the  $(N - 1)^{\text{th}}$  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.’ See Schmidhuber, “Deep learning in neural networks: An overview.”

<sup>94</sup> “Moore’s Law” was an observation 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. See Tom Simonite, “Moore’s Law Is Dead. Now What?” MIT Technology Review, 13 May 2016, <https://www.technologyreview.com/2016/05/13/245938/moores-law-is-dead-now-what/>.

neural networks (DLNNs) was finally introduced in 2006.<sup>95</sup> That year, AI researchers also began exploiting the massively parallel computing powers of graphical processing units (GPUs) to speed up learning even further.<sup>96</sup> The result of both developments was that a large—and increasing—number of “narrow AI” problems showed signs of having been effectively “solved” (see discussion in next section, “**Emerging AI Themes and Issues**”).

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), all of “narrow AI’s” successes share two basic characteristics to date:

1. *They map fairly simple inputs to outputs* (though the meaning of “simple” may be subject to interpretation). For example, an image (as an 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 3-D 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 timescales 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.

Brynjolfsson and Mitchell give a more complete rubric to see if a task can be solved with ML:<sup>97</sup>

- *Data needed to complete task (inputs) and outputs specified in machine-readable format*
- *Task information is recorded or recordable by computer*
- *Task does not require a wide range of complex outputs (mental and/or physical)*
- *Task feedback (from outputs) is immediate or available through plentiful historical data*
- *The task output is error tolerant*

---

<sup>95</sup> Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation* 18, no. 7 (July 2006).

<sup>96</sup> Speedups close to *two orders of magnitude* have been reported: Rajat Raina, Anand Madhavan, and Andrew Y. Ng, “Large-scale deep unsupervised learning using graphics processors,” *Proceedings of the 26th Annual International Conference on Machine Learning*, Association for Computing Machinery, 2009.

<sup>97</sup> Erik Brynjolfsson and Tom Mitchell, “What can machine learning do? Workforce implications,” *Science* 358, 22 Dec 2017, <https://science.sciencemag.org/content/358/6370/1530>.

- *It is not important that outputs are perceived to come from a human*
- *Task does not require complex, abstract reasoning*
- *Task principally concerned with matching data to concepts, labels, predictions, or actions*
- *Task does not require detailed, wide-ranging conversational interaction with a person*
- *Task is highly routine and repeated frequently*
- *Task is describable with rules*
- *There is no need to explain decisions during task execution*
- *Task convertible to answering questions, ranking, predicting, or grouping objects*
- *Long term planning is not required to successfully complete the task*
- *Task requires working with text data or might require working with text in the future*
- *Task requires working with image/video data now or in the future*
- *Task requires working with speech data or might require speech data in the future*
- *The task requires working with other types of machine-readable data*
- *Many components of the task can be completed in a second or less*
- *Each task instance, completion, or execution similar to other instances in how it is done*
- *Task actions to be completed in a very specific order, and practicing to get better is easy.*

## **AI as a categorical taxonomy of algorithms**

As alluded to at the beginning of this section, one half of our oft-repeated mantra to be mindful of AI's "devil in the details" consists of appreciating that "AI" does not refer to any *one* algorithm, but is merely a label that sits atop a large categorical taxonomy of approaches, functions, and methods. This simple *truth* behind what "AI really consists of" is almost always ignored (at least by publications targeting nontechnical audiences). Although it is not necessary for stakeholders and policy-makers to fully comprehend the technical meaning of each and every entry in AI's taxonomy (*that* will be presented shortly), it is vital that they appreciate the fact that *an irreducibly complex taxonomy exists at all*. Why *vital*? Because without having even a rudimentary understanding of what AI's taxonomy is (and, yes, it is messy!), the process of adjudicating between, say, ML method *X* over ML method *Y* (for a problem *P*, in context *C*, under assumptions *A*, and subject to limitations *L*)—or deciding which of a dozen or more proposed AI-based S&T portfolios to fund—is prone to be ill-conceived, at best, and fundamentally flawed, at worst.

It is beyond the scope of this paper to provide anything more than a cursory summary of AI's many interlocked and rapidly evolving parts (for which there are ample pedagogical resources).<sup>98</sup> For our purposes, it suffices to emphasize the need to

- **Acknowledge that AI consists of a veritable zoo of overlapping and mutually supporting techniques**, the lack of a universal appreciation of which often muddies and obscures the “dialogues” that need to take place among different stakeholders with varying levels of technical sophistication.

“AI” is all too often used in conversation and policy deliberations—if only implicitly—as if it referred to some well-defined body of “ready to use” methods and algorithms. While development frameworks that provide a palette of basic ML functions exist,<sup>99</sup> “AI” is a vastly broader field of endeavor. Hence, we must also

- **Understand how the individual components of this “zoo” are all related**, whether in terms of the kinds of problems they are designed to help “solve,” in terms of learning style, or in terms of form and function.

The present discussion is another piece of our “stage setting” narrative that culminates in a framework to help foster “informed” stakeholder dialogues (between military decision- and policy-makers, one side, and AI/ML researchers, developers, and programmers, on the other). Before presenting a full example of an AI taxonomy, we look at a “simple” top-level view of a basic taxonomy.

Figure 6 is a screenshot of one of the slides used during a recording of an episode of CNA’s “AI with AI” podcast series that included a panel discussion on the subject, “*What is AI?*” (and, as such, was designed for a “general” audience).<sup>100</sup>

---

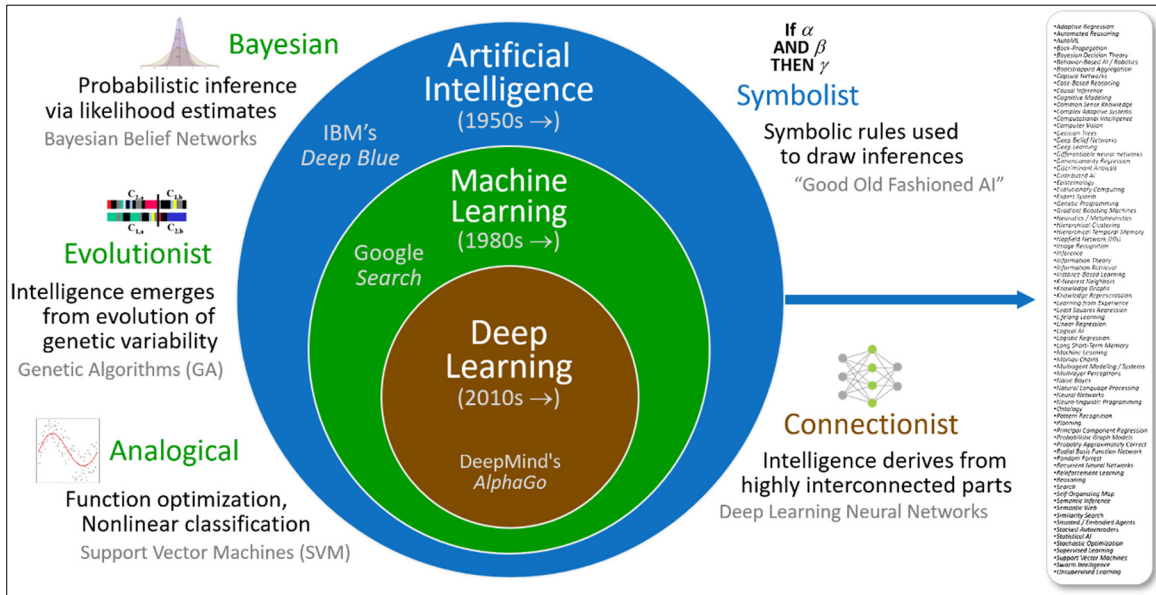
<sup>98</sup> Russell and Norvig, *Artificial Intelligence: A Modern Approach*; Jeremy Watt, Reza Borhani, and Aggelos K. Katsaggelos, *Machine Learning Refined: Foundations, Algorithms, and Applications*, 2nd ed., New York: Cambridge University Press, 2020; John Krohn, Grant Beyleveld, and Aglaé Bassens, *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*, Boston: Addison-Wesley, 2020.

<sup>99</sup> A comparison between two main frameworks (*PyTorch* and *TensorFlow*) is given by Horace He, “The State of Machine Learning Frameworks in 2019,” 10 Oct. 2019, <https://thegradient.pub/state-of-ml-frameworks-2019-pytorch-dominates-research-tensorflow-dominates-industry/>.

<sup>100</sup> The slide-deck from which figure is taken was handed out to an audience assembled at CNA in January 2019 to participate in a *video* recording of the “AI with AI” podcast’s hosts discussing the SOTA in “What is AI? A Panel Discussion on the Opportunities and Challenges Presented by AI,” 22 Jan. 2019, [https://www.cna.org/CNA\\_Files/PDF/AI-Vidcast-slides.pdf](https://www.cna.org/CNA_Files/PDF/AI-Vidcast-slides.pdf).



Figure 6. A schematic illustration of the "Five Tribes" of AI



Source: "AI with AI."

There are three salient takeaways here:

1. AI, having started in the 1950s, includes (but is more general than) the machine learning focus that ensued in the 1980s, which in turn is more general than the deep learning techniques currently in fashion.
2. Among the many equally valid ways of segmenting the AI field as a whole, one illustrative decomposition proceeds along lines of "Five Tribes"<sup>101</sup>—*Bayesian approaches*, which rely on probabilistic inferences vis likelihood estimates; *Symbolist approaches*, which are throwbacks to the 1980s' and 1990s' "Good Old-Fashioned AI," and rely on logic and symbolic rules; *Analogical approaches* that consist largely of classical function optimization and nonlinear classification methods; *Evolutionist approaches*, which are inspired by biological evolution (wherein "solutions" to a problem—even intelligence itself—are "grown" or evolved using basic natural evolutionary processes such as recombination, crossover, and mutation); and *Connectionist methods*, which is another label for what is arguably today's singularly most popular class of deep neural-network learning techniques.
3. Drilling down even a single level from AI's top-most "Five Tribes" decomposition reveals a (deliberately) too-small-to-comfortably-read litany of specific methods, functions, and algorithms (referring to the list that appears on the right-hand side of

<sup>101</sup> The "Five Tribes" decomposition is borrowed from Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books, 2015.

Figure 6). Remarkably, this long list contains but a *few* of AI’s extant methods. Since the author did not have time to organize an actual taxonomy at the time of the video podcast recording (in January 2019), individual entries on this list are listed alphabetically, not according to how they are related. The only point this part of the slide was intended to make (at least at the time of the podcast recording) was that this list is *necessarily long*.

Figure 7 shows a top-level view of a taxonomy of ML organized by *functional similarity*, wherein each major cluster of algorithms is accompanied by a visual schematic that serves as a mnemonic reference for what a given category of algorithms is designed to *do*. The mindmap in **Appendix E** contains the full version of this taxonomy, which includes 100-plus specific algorithms and (embedded hot-link) references to primary and secondary reference sources.

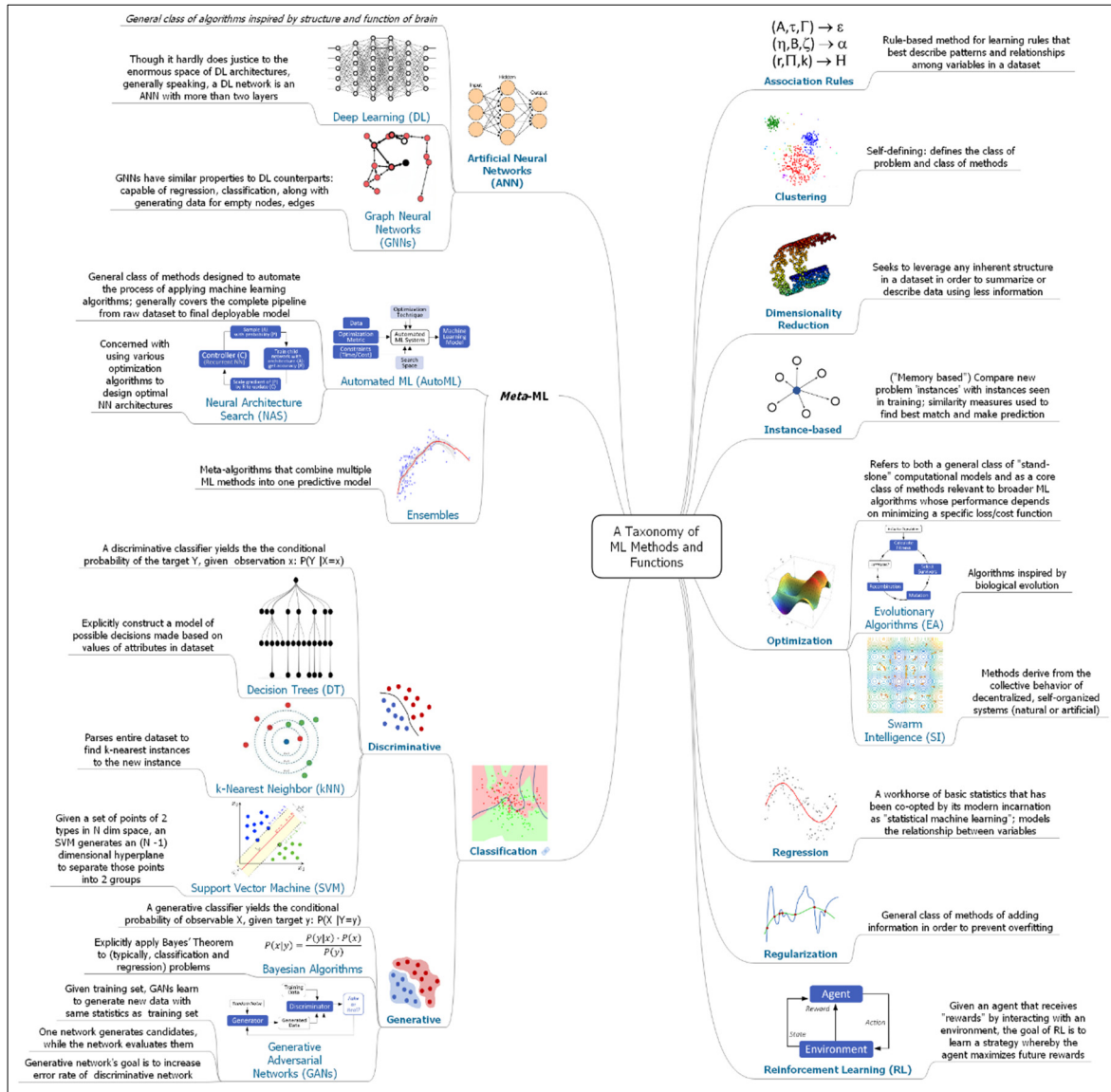
### **AI as a field of scientific discovery**

A complementary view of AI as a taxonomy of methods and algorithms (discussed in the previous section) is that of AI as a “field of scientific discovery.” By this, we mean that, just as for other scientific endeavors, AI may be viewed as both (1) a *problem-driven, process-oriented activity* that has thus far stretched over close to seven decades, and (2) an intensely detail-filled, nested *pipeline of doing basic research* that consists of collecting and preparing data, adopting existing methods to new problems and challenges, and developing deployable systems. Both views represent important additional stage-setting “ingredients” for our narrative—namely, they provide military stakeholders and policy-makers a deeper context in which to understand what “AI really is,” thus enabling them to engage in a more *mutually meaningful dialogue* with AI R&D communities.

### **The “adjacent possible” in an AI/Neural LEGO World**

Figure 8 shows one way to illustrate AI as a “problem-driven, process-oriented activity.” It contains a mindmap that displays a (partial) timeline of various neural network designs and architectures throughout the last sixty years or so. (A more complete—and much easier to read higher-resolution—version that also contains embedded hot-link references to primary source material appears in **Appendix G**.)

Figure 7. A visual taxonomy of ML methods and functions (top-level view only)



Source: CNA.

The timeline starts in 1958 with Rosenblatt’s *perceptron*<sup>102</sup> in the bottom-left of the figure, and ends at top-right with *Google’s* introduction of “Automated ML” (or “AutoML”) techniques in 2017.<sup>103</sup> The short descriptions of each milestone are highlighted by (1) *color-coded diagrams* that show the basic elements of a given design; (2) the fundamental *problem* that a given design was either motivated by and/or developed as an approach to “solving” (highlighted in **red**);

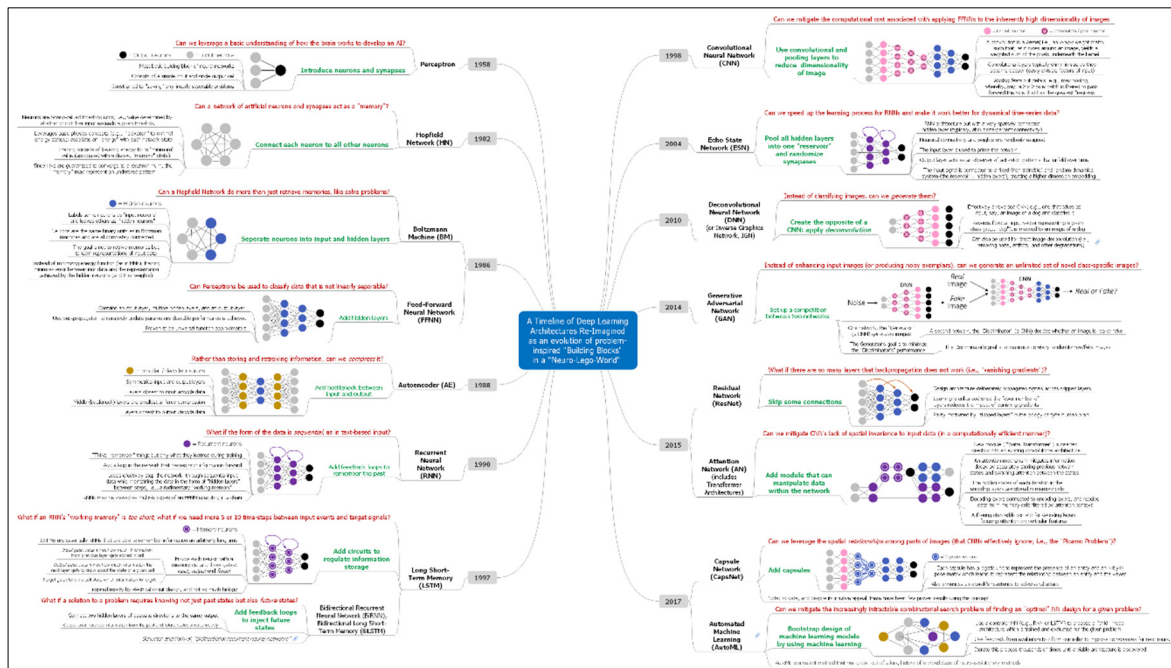
<sup>102</sup> Rosenblatt, “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.”

<sup>103</sup> Quoc Le and Barret Zoph, “Using Machine Learning to Explore Neural Network Architecture,” Google AI Blog, 17 May 2017. <https://ai.googleblog.com/2017/05/using-machine-learning-to-explore.html>.

and (3) the most important *innovation* that was introduced by a given design (highlighted in green).

For example, the perceptron, which introduced artificial neurons and synapses (that remain the basic building blocks of modern-day deep neural networks, albeit with a litany of newly developed fine-tuned features), was a direct “answer” to the question, “Can we leverage a basic understanding of how the brain works to develop an AI?” Implicit in this first answer was the concomitant introduction of two different kinds of neurons: *input neurons* and *output neurons* (highlighted in grey and black, respectively, in the figure). The perceptron was undeniably an innovative design, but it was limited to being able to only “solve” (i.e., classify) the special class of linearly separable problems (as discussed earlier).

Figure 8. Mindmap of milestone developments of NN designs and architectures



Source: CNA.

Among a host of other intuitive follow-on questions that could (and were) asked about Rosenblatt’s general approach is the question, “Rather than ‘solve’ a linearly separable problem, can an artificial neural network serve as working ‘memory’?” The answer to which appeared in the form of the *Hopfield network* (HN), whose innovation involved connecting each neuron to all other neurons.<sup>104</sup> The training of an HN consists of lowering energy to its minimum value (associated with a desired “memory” state).

<sup>104</sup> J. J. Hopfield, “Neural Networks and Physical Systems with Emergent Collective Computational Abilities,” *Proceedings of the National Academy of Sciences* 79 (1982): 2554–2558, <https://doi.org/10.1073/pnas.79.8.2554>.

Soon after, two other (in hindsight, “obvious”) questions were asked: (1) “Can a *Hopfield* network do more than just retrieve memories, like solve problems?” and (2) “Can perceptrons be used to classify data that are not linearly separable?” The solutions to which arrived in 1986: As an answer to the first question, the *Boltzmann machine* (BM) introduced a new class of *hidden neurons* into an HN’s architecture (the goal of which is not to retrieve memories but to learn representations of input data). As an answer to the second question, *multi-layered feed-forward neural networks* added hidden layers into the basic perceptron design (so that the entire architecture a full suite of input, hidden, and output neurons), along with a new training algorithm (“backpropagation”) to minimize the error between input data and the output representation.

Other approaches, designs, neuronal and synaptic types, and architectures all followed in succession, punctuated by fits and starts, periods of stasis followed by rapid development, and a seemingly unyielding propensity for ever-increasing complexity: compare the “simple” design of the erstwhile perceptron or even circa 1990 *recurrent neural networks*, which added feedback loops within the hidden neuron layers to help simpler networks “remember” the past, to the novel (and not always amenable to human understanding) circa 2017 neural network architectures that are bootstrapped using machine learning techniques. Indeed, the evolutionary path thus far taken by AI as a whole, and in the space of “NN design architectures,” in particular, mimics that of other complex technologies’ progress.

It has been suggested (by Brian Arthur,<sup>105</sup> an economist and complexity theorist at the Santa Fe Institute, and other researchers<sup>106</sup>) that the dynamic mechanisms driving technological innovation are similar to those at play in biology—namely, a mix of Darwinian evolution and a combination of earlier (forms of) technologies. If this analogy is valid, the mindmap in Figure 8 may be viewed as a slice through the space of the “adjacent possible” (a term introduced by another complexity theorist, Stuart Kauffman, to describe how biological systems morph into complex systems).<sup>107</sup>

As Steven Johnson explains Kauffman’s idea in his book, *Where Good Ideas Come From*,<sup>108</sup>

In human culture, we like to think of breakthrough ideas as sudden accelerations on the timeline, where a genius jumps ahead fifty years and invents something that normal minds, trapped in the present moment, couldn’t possibly have come up with. But the truth is that technological (and scientific) advances rarely break out of the adjacent possible; the history of cultural progress is, almost without exception, a story of one door leading to another door, exploring the palace one room at a time.

---

<sup>105</sup> W. Brian Arthur, *The Nature of Technology*, New York: Free Press, 2009.

<sup>106</sup> John Ziman, ed., *Technological Innovation as an Evolutionary Process*, New York: Cambridge Univ. Press, 2008.

<sup>107</sup> Stuart Kauffman, “Innovation and the Evolution of the Economic Web,” *Entropy* 21, no. 5 (September 2019).

<sup>108</sup> Steven Johnson, *Where Good Ideas Come from: The Seven Patterns of Innovation*, NY: Penguin Books, 2011.

Gutenberg invented the printing press by creatively adopting and combining pre-existing elements within his available space of the “adjacent possible” (e.g., ink, paper, and movable type). If *Gutenberg* had not done it, someone else assuredly would have. As another example, the inventors of the transistor intuitively understood that the “time was ripe” for the transistor to enter the space of the adjacent possible (following fundamental discoveries of quantum mechanics made during the 1920s through the 1940s): “There was little doubt, even by the transistor’s inventors, that if [William] Shockley’s team at Bell Labs had not gotten to the transistor first, someone else in the United States or in Europe would have soon after.”<sup>109</sup> And, as one winner of the Nobel Prize in Physics observed: You do not do things “until a background knowledge is built up to a place where it’s almost impossible not to see the new thing, and it often happens that the new step is done contemporaneously in two different places in the world, independently.”<sup>110</sup>

Our narrative embraces this basic idea by offering it as a conceptual playground in which to imagine possible futurescapes and AI technologies that all derive and evolve within a vast space of the adjacent possible. We call it an “AI/Neural LEGO World” (AINLW) after Kauffman, who used LEGO blocks as an intuitive way of understanding the adjacent possible *algorithmically*. One starts with a vast collection of LEGO blocks arranged on a central circle (call it “ring 0”) surrounded by a vast set of increasingly larger concentric circles: ring 1 contains all LEGO objects that can be constructed using LEGO blocks in ring 0 by performing a single legal assembly (or “move”). In ring 2, place all objects that can be constructed in two steps; in ring 3, place all objects that be constructed in three steps; and so on. The LEGO structures that currently exist, say, in ring 10, define the adjacent possible of all LEGO structures that can be constructed using any of these structures as the starting point and making a single “move.” Kauffman offers a plausibility argument for how, in realistic complex systems, the adjacent possible is fundamentally non-algorithmic—that is, it is *unprestatable*. Kauffman asks us to think of a screwdriver: *there is no algorithm that can list all possible uses of it!* Taking this analogy a step further, just as AI methods live in and are in principle all constructible in a vast but explorable AINLW space, it behooves the military to develop a set of tools to explore, navigate, discover, and tinker with “operationalizably constructible” designs in selected subsets of this space. The framework that is introduced in the last section of this paper is a step toward this goal.

## The workflow pipeline for doing AI

Figure 9 shows a schematic illustration of AI as a “pipeline of doing basic research.” Just as part of the reason for displaying the hard-to-read list of AI/ML methods on the right-hand side of Figure 6 is to emphasize the fact that this list is *long* (and to help dispel the notion that what AI

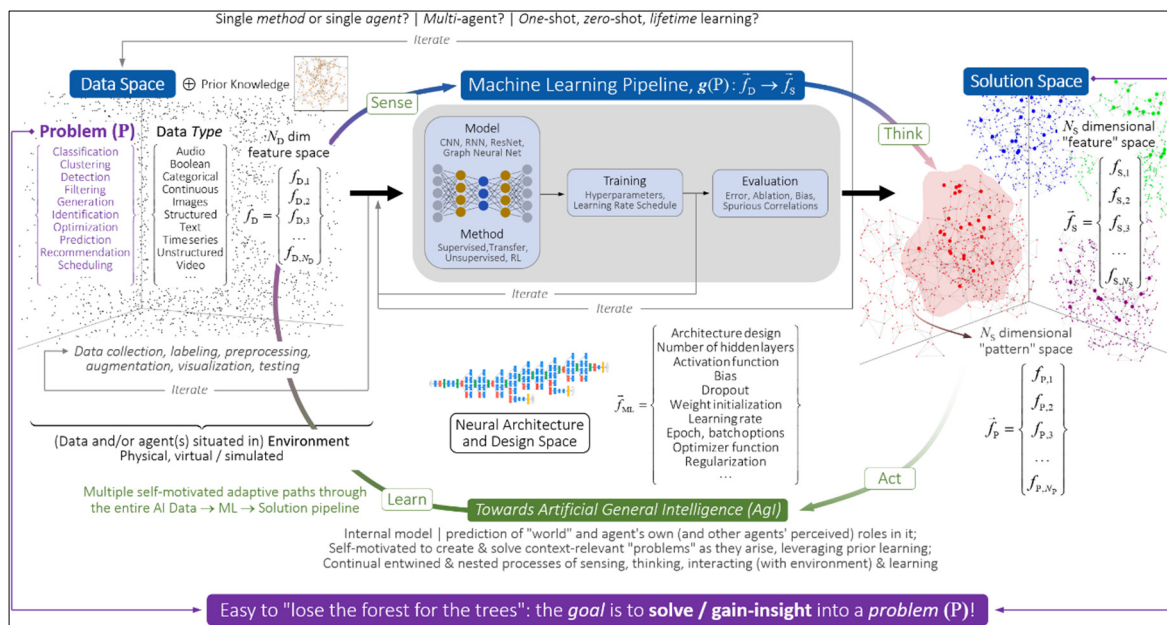
---

<sup>109</sup> Jon Gerner, *The Idea Factory: Bell Labs and the Great Age of American Innovation*, New York: Penguin Books, 2012.

<sup>110</sup> Quoted on p. 204 in Harriet Zuckerman, *Scientific Elite: Nobel Laureates in the United States* New York: Routledge, 2018.

really is can be meaningfully captured in a pithy sentence-long definition), so too the busyness of the workflow that appears in Figure 9 is deliberate, and is intended to convey the *irreducible* “devil is in the details” level of complexity that “doing AI” actually entails.

Figure 9. A schematic illustration of a typical AI/ML development pipeline



Source: CNA.

Reading off the parts of the schematic from left to right, the typical pipeline for developing deep learning applications (on a conceptual level)<sup>111</sup> can be viewed as consisting of a series of nested, entwined activities that take place within three abstract spaces (highlighted in blue): (1) the **data space**, which is the lifeblood of most deep-learning applications and within which multiple steps need to be taken; (2) the **neural architecture and design space**, within which “machine learning” takes place (inside the ML pipeline);<sup>112</sup> and (3) the **solution space**, within which the “output” of a trained system resides (e.g., the space of labels for an ML designed to identify parts of images).

But these simple descriptions belie the complexity that lives underneath. Each of these spaces harbors multiple layers of activity, the existence of which (in typical high-level overviews for general audiences) is either glossed over, at best, or simply ignored, at worst. For example, activities within the data space include collection, labelling, preprocessing, and augmenting (in

<sup>111</sup> We do not have the space here to include a discussion of the equally important *practical* workload aspects of this workflow (e.g., selection and/or stand-alone development of data collection and data preparation tools, ML software frameworks—such as Keras, PyTorch or TensorFlow, and validation and analysis toolkits).

<sup>112</sup> A more detailed discussion of the middle part of this schematic (i.e., the ML pipeline) than appears here is given by Maithra Raghu and Eric Schmidt, “A Survey of Deep Learning for Scientific Discovery,” DeepAI, 26 March 2020, <https://deepai.org/publication/a-survey-of-deep-learning-for-scientific-discovery..>

the event that available data are insufficient). Also implicit in these data-related activities is an additional need to respect environmental and problem-specific conditions and constraints and/or integrate prior knowledge.

Learning-focused steps include selecting a specific deep neural network model (which may require a “meta design” stage and use of AutoML methods to find a suitable architecture), choosing the task and method to train the model, and evaluating the efficacy of the training (which itself includes numerous activities such as validation, analysis, interpretation of hidden representations, and the selection of appropriate ablation methods).

Reaching the “solution space” is not—as commonly viewed—the final step in which an “answer” is revealed (e.g., the ML method outputs “deep fake” video), but rather best viewed as the *first* step in what is almost always a long iterative process in which various steps of the *data processing* → *training* → *validation* cycle are repeated multiple times.

The thin gray backward-flowing arrows (labeled “iterate”) in Figure 9 highlight a few of these iterations: an image-labelling process may not have captured certain categories of interest on the first pass through the data space, requiring another pass; the first pass through the ML-pipeline may underperform on a classification task and require additional passes using supervised learning; or the analysis of the ML-pipeline’s output may reveal that the learning method has overfit on the training data, thus requiring another pass with a shorter training time and/or a different architectural design.

The series of circular flows that pass through the labels *sense*, *think*, *act*, and *learn* (highlighted in **green**), schematically denote one possible (albeit entirely notional) artificial *general* intelligence (AGI) pipeline. We expect this pipeline to consist of the same general steps as those making up the *human*-design process, but to be *self-driven*—that is, to proceed according to internal predictions about the state of the environment in which it is situated (i.e., its “world”) and about both its own and other AI-agents’ perceived roles in it, to be self-motivated to create and solve “problems” as they arise, leveraging and/or adapting context-specific prior knowledge and learning and generally to be continually entwined with ongoing processes of sensing, thinking, interacting (with environment), and learning on its own. We will know when a genuine AGI arises when an “AI” system *decides on its own* to either:

- redefine a problem (e.g., to better align its training regimen with an ultimate goal that is only implicitly stated);
- perceive a need for additional information that it was not originally trained on and proceed to find, collect, process, and integrate new data into its learning cycle; and/or
- “redesign” its own architecture and ML pipeline.

But, even disregarding these notional AGI components (which are discussed here more for reference and completeness rather than as substantive elements of the main narrative), we urge the reader to remember AGI’s underlying *sense* → *think* → *act* → *learn* cycle before



reading the last section (where it will be used to forge connections among AI algorithms, human cognitive abilities, and military applications).

Perhaps the *most* important part of Figure 9 has nothing at all to do with AI (or AGI) per se and has heretofore not yet been explicitly mentioned: the **problem space** (highlighted in purple). Reflecting once again on the *busyness* of Figure 9, as faithful to the practical reality of the difficulty of developing AI systems as the schematic in this figure is, it is easy to lose the forest for the trees: **the ultimate goal is to solve or otherwise gain useful insight into a problem!** Certainly, AI *may* offer a solution, but it is not necessarily the only, or “best,” method of finding a practicable—or, in a military context, *operationalizable*—one to use. This basic observation assumes a particular urgency in light of the rising specter of an impending stall-in-progress of ML methods (if not an outright “third AI winter”), discussed in a later section.

For DOD, the question to ask is not necessarily, “Which new AI approach do we invest in the development of?” Rather, it is, “Which existing method best ‘solves’ our problem?” As we will see, depending on the problem, AI is not always the wisest choice. And, if a given ML method *does* offer the “best” solution, it may not entail as heavy an investment in cutting-edge technology as is frequently assumed: yesterday’s tried-and-true algorithms often outperform their modern-day brethren. Thus, DOD’s *meta*-problem is to adjudicate among investments in developing and applying new AI approaches and adapting or adopting existing methods.

## AI’s perennially persistent fundamental gaps, challenges, and limitations

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 (e.g., backgammon, chess, Go, Othello, poker, Scrabble, visual pattern recognition, etc.), there are situations when an AI’s “solution” is *surprising* and/or *blatantly wrong*.

Although 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 Go match he lost in 2016), 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” (i.e., the *explainability problem*).<sup>113</sup> 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 (during, say, the testing and evaluation phase of

---

<sup>113</sup> Arun Das and Paul Rad, “Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey,” 23 June 2020, arXiv:2006.11371.

DOD’s acquisition process) that the autonomous system being developed will not perform “surprising” (i.e., unanticipated) actions during a mission?

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 counterintuitive 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 (DLNN) to classify the image as something entirely different (such as the well-known “panda” example),<sup>114</sup> 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 percent confidence (e.g., labeling with certainty that white noise static is a lion).<sup>115</sup>

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.<sup>116</sup> 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.

The following list contains a long (but far from exhaustive!) list of as-yet-unsolved technical challenges and methodological limitations for circa 2020 AI and ML techniques:

- Intensely data hungry—require vast relevant datasets with labeled exemplars
- “Devil in the details” complexity of AI and ML development pipelines is highly nontrivial

---

<sup>114</sup> Such as the well-known example 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. See Ian Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *Proceedings of the International Conference on Learning Representations (ICLR)*, 20 March 2015: <https://arxiv.org/pdf/1412.6572v3.pdf>.

<sup>115</sup> Anh Nguyen, Jason Yosinski, and Jeff Clune, “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images,” *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, [http://www.evolvingai.org/files/DNNsEasilyFooled\\_cvpr15.pdf](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. See also A. Ilyas, et al., “Adversarial Examples Are Not Bugs, They Are Features,” 12 Aug 2019. arXiv:1905.02175.

<sup>116</sup> Mike James, “The Flaw Lurking In Every Deep Neural Net,” *I Programmer*, 27 May 2014, <https://www.i-programmer.info/news/105-artificial-intelligence/7352-the-flaw-lurking-in-every-deep-neural-net.html>.

- Inherently brittle—mature methods typically narrowly focused; do poorly outside constraints
- Irreducibly fragile ("Manifold hypothesis")—vulnerable to attack and/or exploitation
- Basic research concerns (e.g., *replicability* and *reproducibility*)
- Inherently opaque—*explainability*, *predictability*, and *understandability*
- Fundamental limits on ability to anticipate emergent behaviors, particularly for AI-infused complex systems-of-systems
- Not well integrated with prior knowledge
- Lack of common sense and limited “understanding” of context (that humans take for granted)
- Limited capacity for transfer (to other problems/domains)
- Does not easily distinguish causation from correlation
- Struggles with open-ended inference and general reasoning capacity about environment
- Difficulty with exploration games with sparse rewards (reinforcement learning, or RL, methods)
- Lives best in *static* universes
- Only nascent development of *meta-learning* and *lifelong learning*
- Difficulty in maintaining human-AI goal and task alignment in dynamic environments
- Deeply prone to the “hype machine”

A mindmap that contains a far more complete taxonomy of AI’s gaps, challenges, and limitations appears in **Appendix F**. The section, “**New AI ‘Challenges’**” highlights more recent examples (that have appeared since 2017). Cogent discussions of AI’s technical challenges appear in recent books by Broussard,<sup>117</sup> Marcus and Davis,<sup>118</sup> and Mitchell.<sup>119</sup>

---

<sup>117</sup> Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge, MA: MIT Press, 2018.

<sup>118</sup> Gary Marcus and Ernest Davis, *Rebooting AI: Building Artificial Intelligence We Can Trust*, New York: Vintage, 2019.

<sup>119</sup> Melanie Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*, New York: Farrar, Straus, and Giroux, 2019.

[This page intentionally left blank]

# Emerging AI Themes and Issues

---

## Recent trends

The exponential growth in AI in the academic and commercial research communities discussed in the 2017/AI paper continues unabated; indeed, there are clear signs that it is accelerating. Not surprisingly, keeping abreast of even a small subset of the latest AI/ML-related research activities, publications, and conference proceedings is becoming increasingly difficult, if not impossible. Since the goal of this paper is to provide insights into emerging trends and issues, and not to conduct an exhaustive survey (for which there are plenty of available resources),<sup>120</sup> we will limit our exposition to a few basic stage-setting statistics and observations.

Figure 10 gives the broadest possible overview of how general interest in AI has evolved over time (2009–present) and is culled from a Google Trends search using five search phrases:<sup>121</sup>

- *Artificial intelligence,*
- *Data science,*
- *Deep learning,*
- *Machine learning,* and
- *Statistical analysis.*

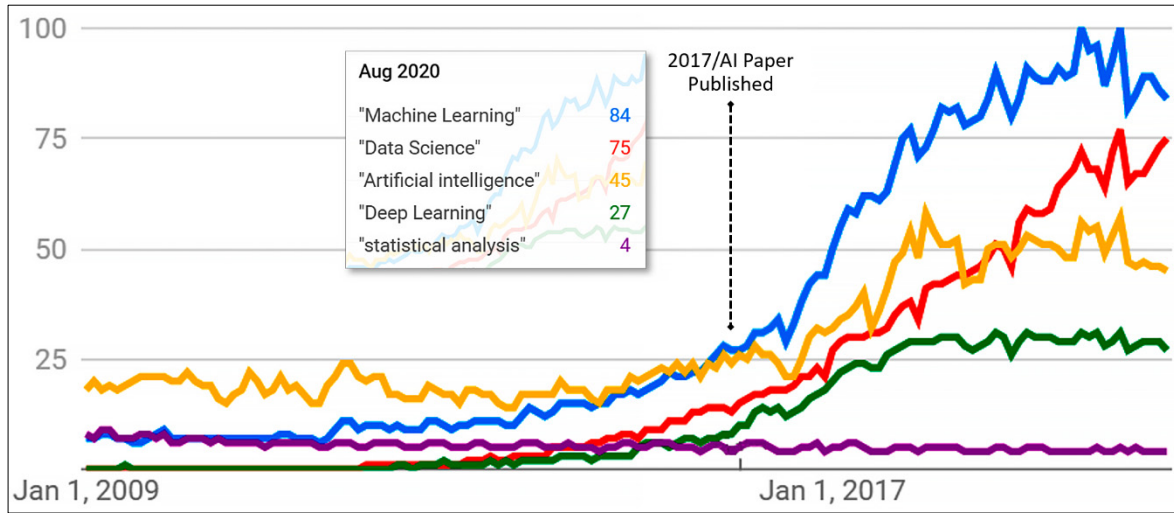
The main takeaway is a simple confirmation that worldwide interest in AI continues to grow. Although it is tempting to draw some additional inferences (e.g., interest in conventional statistical methods, highlighted in **purple**, appears to be waning in comparison to AI/ML, and “artificial intelligence” is less trendy than “data science,” since both assertions are demonstrably true for the specific search phrases used to generate), Figure 10’s data are far too coarse to use to reveal any fine structure in the statistics.

---

<sup>120</sup> A survey-of-surveys website contains links to more than 100 recent technical surveys on various AI and ML methods: <https://github.com/NiuTrans/ABigSurvey>. Other references include Stanford University’s Human-Centered AI Center’s Global AI Vibrancy Tool (<http://vibrancy.aiindex.org/>) and annual “AI Index” reports (2019 version: <https://hai.stanford.edu/research/ai-index-2019>); the online repository, <https://paperswithcode.com/>; arXiv, a real-time monitoring tool of AI/ML preprints: <https://www.nesta.org.uk/blog/arxlive/>; and Dimensions.ai’s open-resource repository that allows interactive search of a large number of preprint servers and open-access journals (<https://app.dimensions.ai/discover/publication>). A master list of links for all major AI/ML-related international conferences (each of which typically provides its own master list of current and past proceedings) is available at <https://developer.att.com/blog/ai-conferences>.

<sup>121</sup> According to the Google Trends website, the numbers on the y-axes (between 0 and 100) “represent search interest relative to the highest point on the chart for the given region and time.” Thus, 100 is the peak popularity for a given term, and 50 means that the term is half as popular; see <https://trends.google.com/trends>.

Figure 10. Google Trends statistics for five AI-related key phrases



Source: CNA.

We can take a deeper dive into recent trends by looking at timelines of the number of AI/ML-related research papers that have been posted to the preprint website ArXiv.<sup>122</sup> ArXiv is a popular website used by the technical research community to post and share work that often later appears in journals (after passing peer review). ArXiv does not just cater to artificial intelligence. The top level of its category taxonomy includes eight disciplines:<sup>123</sup> computer science, economics, electrical engineering and systems science, mathematics, physics, quantitative biology, quantitative finance, and statistics. "Artificial intelligence" (CS.AI) is a stand-alone category and is the first major branch within computer science (CS). Overall, CS includes 39 additional topics, many of which overlap with AI methods and technologies (e.g., computer vision and pattern recognition, machine learning, multi-agent systems, robotics, and symbolic computation).

What makes ArXiv particularly appealing as a dataset from which to glean insight into research trends is the fact that it is extensively used by AI/ML researchers: a large fraction of papers presented at major conferences are available on ArXiv. Moreover, almost all papers from leading AI R&D labs (DeepMind, OpenAI, Facebook AI, etc.) are also available from ArXiv. Thus tracking papers that discuss specific subjects (and combination of topics) gives important clues about how ideas are evolving and diffusing over time.<sup>124</sup> Still, an overall caveat must be applied to all of the observations made in this section: as large an archive as ArXiv unquestionably is,

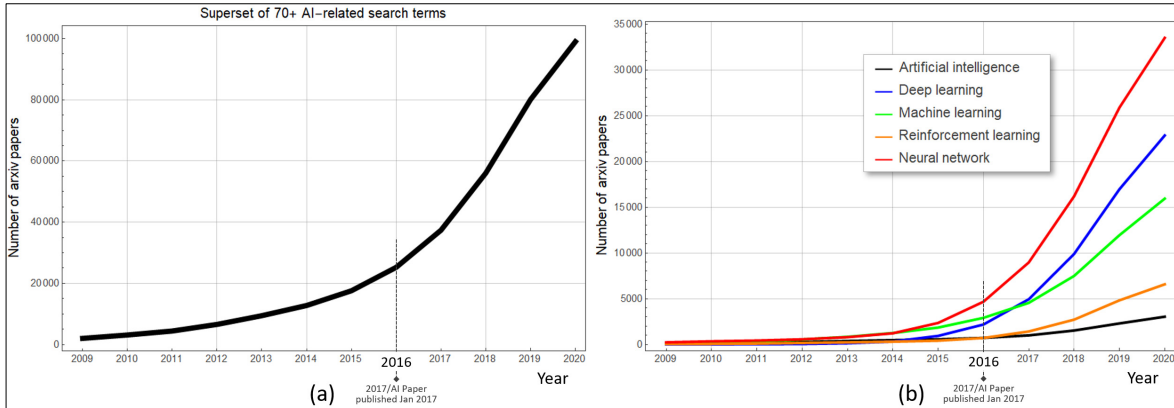
<sup>122</sup> Arxiv homepage: <https://arxiv.org/>.

<sup>123</sup> [https://arxiv.org/category\\_taxonomy](https://arxiv.org/category_taxonomy).

<sup>124</sup> ArXiv announced on 5 August 2020 that it is releasing a free, open pipeline on Kaggle to its entire machine-readable database, a repository of more than 1.7 million papers: <https://blogs.cornell.edu/arxiv/2020/08/05/leveraging-machine-learning-to-fuel-new-discoveries-with-the-arxiv-dataset/>.

the papers added to it on a daily basis represent only a fraction of the work being done in AI/ML at any given moment.

Figure 11. Number of AI-related papers posted to ArXiv.CS between 2009 and 2020



Source: CNA.

Bearing this caveat in mind, Figure 11a shows the total number of AI-related papers posted to ArXiv between the years 2009 and 2020. ArXiv’s search engine was used to tally all documents in the CS category that contained any of 70-plus specific AI-related terms and phrases in either their title or abstract.<sup>125</sup> Where there were slightly over 25,000 total papers before 2017/AI was published (January 2017), that number has since almost *quadrupled!* (~99,138, as of 31 August 2020).

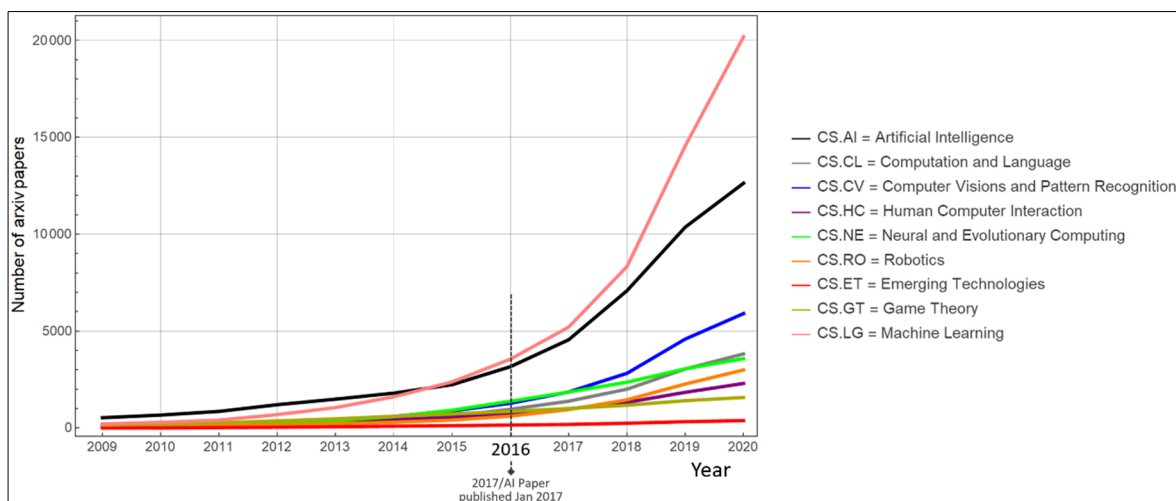
Figure 11b shows a finer breakdown according to five specific top-level phrases: *artificial intelligence*, *deep learning*, *machine learning*, *reinforcement learning*, and *neural network*. As for the trends shown in Figure 10, one ought not read too much into absolute numbers. Nonetheless, it is tempting to speculate that the relative paucity in numbers for “artificial intelligence” compared to other phrases is due to its diminution in specificity over time (as newer and finer-grained methodologies have become increasingly more familiar). The slight dip that is evident at the tail end of each of the curves in both Figure 11a and Figure 11b is due to the fact that the 2020 data is incomplete (i.e., the tally includes only those papers posted through 31 August 2020).

Figure 12 shows the total number of papers posted to selected AI/ML-related *branches* of the computer science (CS) category in ArXiv’s taxonomy. The difference between the tallies that appear in this figure compared with those in Figure 11 is that the main subjects are specified by the author(s), rather than searching for phrases in titles and abstracts; that is, the numbers

<sup>125</sup> Search phrases include those that are both obvious (e.g., “artificial intelligence,” “machine learning,” and “neural network”) and highly specific (e.g., “hidden Markov model,” “long short-term memory,” and “k-means clustering”). The goal was to get a broadly representative set of AI/ML-related papers. ArXiv’s advanced search page was used for tallying the statistics (counted since 2004): <https://arxiv.org/search/advanced>.

directly reflect *author*-defined category placements at the time a paper was originally posted. It is evident that all nine branches show dramatic increases in volume (over 2.6× increase between 2016 and 2020 for neural and evolutionary computing, 4.6× for computer vision, and more than 5.7× for machine learning). Even “emerging technologies” (sitting at the bottom of Figure 12 in terms of absolute numbers) has shown a 2.5× increase in total numbers.

Figure 12. Number of papers posted to selected AI/ML-related branches of the ArXiv.CS



Source: CNA.

Figure 13 shows the results of a search using some specific phrases that only started to appear (and/or increase in frequency of appearance) in papers posted to ArXiv around 2017. For example, Figure 13a shows that exactly *zero* papers whose title or abstract contained the phrase “neural architecture search” (which today, circa 2020, is an intensely active area of research) were posted prior to 2017. As of 31 August 2020, 427 papers have been published. Similarly, “capsule networks” (capsules are groups of artificial neurons whose output represents different properties of the same entity) were introduced only in October 2017, but have so far garnered a relatively lukewarm reception.<sup>126</sup>

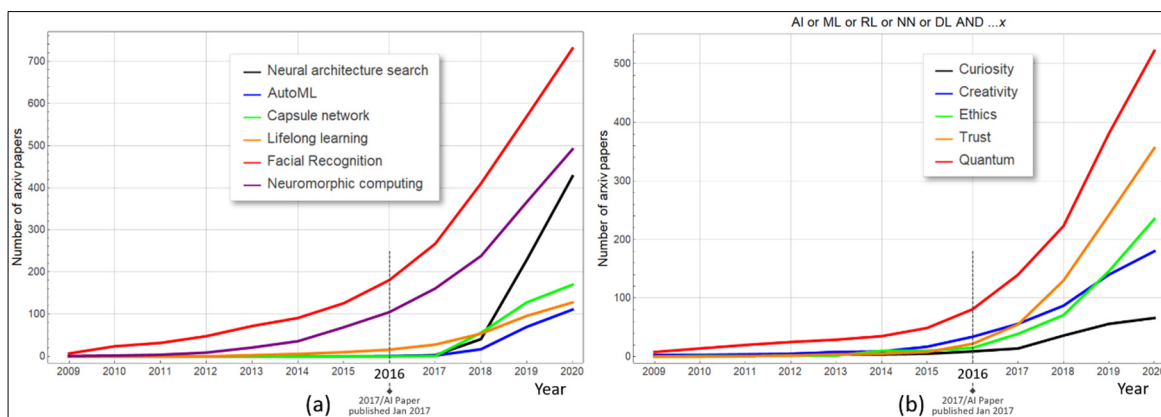
Figure 13b shows the results of a search using the string “artificial intelligence” *or* “machine learning” *or* “reinforcement learning” *or* “neural network” *or* “deep learning” *and* *x*, where *x* is (in each separate case) either *curiosity*, *creativity*, *ethics*, *trust*, *or quantum*. These extra search terms were chosen deliberately to see if ArXiv’s paper post statistics lend credence to certain recent research trends that CNA’s “AI with AI” podcast discussions have only anecdotally

<sup>126</sup> Sara Sabiour, Nicholas Frosst, and Geoffrey E. Hinton, “Dynamic Routing Between Capsules,” 31st Conference on Neural Information Processing Systems (NIPS 2017), <https://arxiv.org/pdf/1710.09829v1.pdf>. Co-author Geoffrey Hinton was awarded the prestigious Turing Award in 2018 and is widely regarded as one of the founders of AI research. Because of this, when the “capsules” papers appeared, the idea was touted as “breakthrough.” Three years later the extent to which it represents a fundamental advance remains unclear. This is an early “microcosm” of one of the leitmotifs of the main narrative of this paper.



hinted at (these trends are discussed in more detail in a later section). In each case, including that of the increasingly important role that “quantum” computation plays as quantum computer technology slowly matures (which, as a *concept*, dates back more than a decade),<sup>127</sup> research-level interest in each of these AI/ML-related “themes” has seen a marked increase only in the last few years.

Figure 13. Number of AI-related papers posted to ArXiv.CS that satisfy specific search phrases



Source: CNA.

For a still deeper dive into what can be gleaned from papers posted to ArXiv, we turn to a recent survey conducted by *MIT Technology Review*.<sup>128</sup> The survey is based on analyzing the abstracts of more than 16,000 papers in the “artificial intelligence” branch of ArXiv’s computer science category (i.e., CS.AI), and includes all papers dating back to 1993 (i.e., the first year that section appeared in ArXiv’s taxonomy). Major historical trends—many of which are discussed in the 2017/AI paper and others that will be touched on in later sections of this paper—are easily discerned from the data: for example, the shift away from knowledge-based systems (i.e., “logic,” “constraint,” and “rule”) to concepts relating to machine learning (e.g., “data,” “network,” and “learning”) during the late 1990s and early 2000s, and the rise in the popularity of neural networks starting in the early to mid-2010s (which took place on the heels of the landmark performance of AlexNet,<sup>129</sup> a deep neural network that took first place in 2012 for Stanford University’s annual “ImageNet Challenge,” since AlexNet was the only neural-network-based entry). More recent trends include the rise in papers that include “reinforcement learning” (which is consistent with our own analysis; see Figure 11b), and is to

<sup>127</sup> *Quantum Computing: Progress and Prospects*, National Academies Press, 2019.

<sup>128</sup> Karen Hao, “We analyzed 16,625 papers to figure out where AI is headed next,” *MIT Technology Review*, 25 Jan. 2019, <https://www.technologyreview.com/2019/01/25/1436/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/>.

<sup>129</sup> A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Communications of the ACM*, May 2017, <https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.

be intuitively expected as a reaction to AlphaGo’s milestone performance against a human Go champion in 2016.

The key takeaway of *MIT Technology Review’s* survey—and one of the leitmotifs of this paper, already touched on in the short history of AI that appears in the previous section (see “**So, what is AI, really?**”)—is that the history of AI is replete with competition and overlaps (in methodological approaches) and is propelled by a drumbeat of progress punctuated by starts, stalls, and sometimes outright halts in research. Every preceding decade was effectively defined by its own unique prevailing zeitgeist:<sup>130</sup> the 1950s and 1960s were the decades of the (early, perceptron-based) “neural networks”; the 1970s introduced an early era of symbolic approaches; knowledge-based expert systems came to the fore in the 1980s; Bayesian networks were dominant in the 1990s; the 2000s ushered in support-vector machines; and (latter-day, “deep”) neural networks arose again in the 2010s. While no one knows what the 2020s will bring, hindsight suggests that we ought not expect them to be any less turbulent or their innovations any less unexpected. The concluding section of this paper—which introduces the “template of a framework” (mentioned in the introduction) to bridge the gap between *understanding* AI and *operationalizing* its military applications—is designed to help DOD mitigate the expected turbulence.

## AI “hits” during 2017–2020

2017/AI’s executive summary begins by asserting that a “notable number of groundbreaking AI-related technology announcements and/or demonstrations took place in 2016” and goes on to list 10 of the (then) “breaking” achievements. The first item on that list highlights a landmark event in 2016 in which AlphaGo defeated 18-time world champion Lee Sedol in the game of Go.<sup>131</sup> Appropriately enough, an updated list that includes milestones achieved between 2017 and August 2020 (see Figure 14 and Figure 15 for a timeline and more examples) begins and ends with updates to AlphaGo:

1. AlphaZero, starting from random play, and using no domain knowledge except for game rules, required only 24 hours to achieve a superhuman level of play in chess, shogi (a Japanese variant of chess), and Go, and defeated a world-champion program in each.<sup>132</sup>
2. Two AIs (one from Microsoft, the other developed by Alibaba) “defeated” humans on the Stanford Question Answering Dataset (SQuAD) for the first time.<sup>133</sup>

---

<sup>130</sup> We do not expand upon the AI/ML methods mentioned here because they are all discussed in later sections.

<sup>131</sup> Koch, “How the Computer Beat the Go Master.”

<sup>132</sup> D. Silver et al., “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm,” 5 Dec 2017, arXiv:1712.01815v1.

<sup>133</sup> Eileen Yu, “Alibaba neural network defeats human in global reading test,” ZDNet, 15 Jan. 2018, <https://www.zdnet.com/article/alibaba-neural-network-defeats-human-in-global-reading-test/>.

3. AI-Feynman "discovered" all 100 test equations from the *Feynman Lectures on Physics*.<sup>134</sup>
4. An ML algorithm used to design chips exceeded human-designed performance.<sup>135</sup>
5. The first AI-developed drug went into clinical trials.<sup>136</sup>
6. An ML algorithm was used to sense people's postures and movement through walls with Wi-Fi.<sup>137</sup>
7. DeepMind's AlphaStar defeated 99.8 percent of human StarCraft II gamers.<sup>138</sup>
8. AI learned to play all 57 Atari video games (in the "Arcade Learning" environment).<sup>139</sup>
9. First AI ("Pluribus") to defeat human professional players in multiplayer game.<sup>140</sup>
10. MuZero matched AlphaZero's superhuman performance without any knowledge of game rules.<sup>141</sup>

One additional AI "hit" appeared just as this paper was being written: on 20 August 2020, an AI "pilot" developed by Heron Systems<sup>142</sup> defeated an Air Force F-16 piloted by a human in a simulated aerial dogfight contest; the final score was 5-0 in the AI's favor.<sup>143</sup> Though this unquestionably qualifies as a "milestone event" in the history of military applications of AI, it is also an instructive microcosm of how *reality is not always what it first seems* when it comes to judging the significance of an AI-related achievement. A typically hype-centric headline reporting the event reads, "AI Claims 'Flawless Victory' Going Undefeated In Digital Dogfight With Human Fighter Pilot."<sup>144</sup>

---

<sup>134</sup> Silviu-Marian Udrescu and Max Tegmark, "AI Feynman: A physics-inspired method for symbolic regression," *Science Advances*, 15 April 2020, <https://advances.sciencemag.org/content/advances/6/16/eaay2631.full.pdf>.

<sup>135</sup> Anna Goldie and A. Mirhoseini, "Placement Optimization with Deep Reinforcement Learning," 18 March 2020, <https://arxiv.org/abs/2003.08445>.

<sup>136</sup> N. Cohen, "Using AI for drug discovery shows speed but draws discussions," *Tech Explore*, 3 Feb, 2020, <https://techxplore.com/news/2020-02-ai-drug-discovery-discussions.html>.

<sup>137</sup> M. Zhao et al., "Through-Wall Human Pose Estimation Using Radio Signals," *Computer Vision and Pattern Recognition (CVPR)*, 2018. <http://rfpose.csail.mit.edu/>.

<sup>138</sup> O. Vinyals et al., "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature* 575, 30 Oct. 2019, <https://www.nature.com/articles/s41586-019-1724-z>.

<sup>139</sup> A. Badia et al., "Agent57: Outperforming the Atari Human Benchmark," 30 March 2020, arXiv:2003.13350.

<sup>140</sup> N. Brown and T. Sandholm, "Superhuman AI for multiplayer poker," *Science* 365, 30 Aug. 2019.

<sup>141</sup> J. Schrittwiser et al., "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model," 19 Nov. 2019, <https://arxiv.org/abs/1911.08265>.

<sup>142</sup> Heron Systems is a small female- and minority-owned company, with offices in Maryland and Virginia, <https://heronsystems.com/>.

<sup>143</sup> Theresa Hitchens, "AI Slays Top F-16 Pilot In DARPA Dogfight Simulation," *Breaking Defense*, 20 Aug. 2020.

<sup>144</sup> Joseph, Trevithick, *The War Zone*, 20 Aug. 2020, <https://www.thedrive.com/the-war-zone/35888/ai-claims-flawless-victory-going-undefeated-in-digital-dogfight-with-human-fighter-pilot>.

Heron Systems' AI's victory took place after a string of earlier events and trials that are part of DARPA's Air Combat Evolution (ACE) program, launched in May 2019.<sup>145</sup> The goal of ACE is to increase warfighter trust in autonomous combat technology. Specific challenges include (1) demonstrating air combat autonomy performance in local behaviors (individual aircraft and team tactical); (2) building and calibrating trust in air combat local behaviors; (3) scaling performance and trust to global behaviors (including multiple heterogeneous aircraft); and (4) building infrastructure for full-scale air combat experimentation. The August 2020 dogfights between AI and human pilots are but one step toward addressing these broader challenges.

So how is *reality not always what it first seems*? Ostensibly, Heron's AI defeated the human F-16 pilot (an Air Force officer—who went by the call sign “Banger”—and instructor at the Air Force's Weapons School at Nellis Air Force Base) and the lopsided 5-0 tally does even not include the fact that “Banger” never managed to get a valid targeting solution and thus was unable to fire a single shot for any of the five dogfights!<sup>146</sup> The “reality” is that the simulated dogfights had at least two significant artificialities that call into question the significance of what was actually accomplished: (1) the (simulated) F-16's weapons loadout was limited to a single notional M61 Vulcan 20mm Gatling gun whose **shots were assumed to automatically hit a target with no misses**; and, most egregiously, (2) **Heron Systems' AI had complete access to the state-space** (i.e., it had perfect situational awareness), whereas the human pilot (who wore a virtual-reality headset that gave him a simulated cockpit-view of combat) had to physically and continually strain his neck and body just to get in position to *see* where Heron's AI F-16 was, not to mention keeping up with the rapid-fire changing data displays in his notional cockpit. Moreover, the Heron AI's winning “tactic,” which consisted of flying directly at its opponent while firing its gun and veering away at the last possible moment—sometimes coming within 100 feet of a midair collision—is not only something a human would not do in practice (human pilots are not allowed to get within 500 feet of each other or to directly face off), but is something a human is also likely never to *choose* to do even in the heat of battle, since doing so may entail colliding with the debris of the destroyed plane and crashing.

If the controlled conditions of the simulation do not allow us to automatically infer that the AI could have beaten a human in real combat, what is the “true” takeaway here? The most salient point is not that Heron's AI bested human performance in an overly constrained environment, but that, given the constraints under which it was required to perform (and for which it was trained), the AI performed exactly as it was designed to perform: *find, within its available search space, a course of action that maximizes probability of success* (defined, in this case, as

---

<sup>145</sup> DARPA's ACE homepage: <https://www.darpa.mil/program/air-combat-evolution>. Heron Systems won the right to engage in the dogfight contest with a human pilot after coming in first in a round-robin tournament (held 18-20 Aug. 2020) that included seven other teams. The dogfight contest took place in JSBSim, a multi-platform open-source flight dynamics model: <https://jsbsim-team.github.io/jsbsim/index.html>.

<sup>146</sup> At the time of this writing (August 2020), details about Heron's methodology were unavailable. In a short Q&A session with three of Heron's AI's developers during the livestream, it was revealed that Heron's AI was trained using reinforcement learning (with the winning agent having been effectively trained for “30 years” in simulated environments); 1:36:45 mark in video of dogfight trials, 18-20 August 2020: <https://youtu.be/NzdhIA2S35w>.

destroying its opponent). Unlike humans, AI is unencumbered by either physiology (Heron's AI held 7 to 8 g's in turns for minutes at a time, and went over 11 g's at one point, which would have made a human pilot pass out) or psychology (Heron's AI was clearly unconstrained by self-preservation instincts). Compared to humans, AI is able to assimilate and find patterns in vastly larger data spaces and make decisions on vastly shorter time scales. *Of course*, it is superhuman! As Col. Daniel "Animal" Javorsek (program manager of DARPA's ACE program) said during the livestream event, the goal of the demonstration was to "increase the confidence of the feasibility of using AI in combat aircraft" and to start toward an eventual "human-machine symbiosis" in which humans "do what they do best" in the cockpit (such as applying broad, contextually relevant, common-sense guidance to mission narratives) and AIs do what *they* do best (such as performing certain sets of difficult but narrowly focused tasks at superhuman levels).<sup>147</sup>

So the real takeaway from Heron's AI's victory over the human F-16 pilot is its behind-the-scenes "reveal," in microcosm, of the myriad as-yet-unresolved basic challenges facing all cutting-edge AI research with potential military applicability: *Is it scalable? Does its performance degrade as artificialities are reduced? How does it perform in a real (not simulated) aircraft? Can its "tactics" (its effective logic) be exploited? Can it adapt to changing environmental conditions and rules of engagement? What will it take to establish a requisite level of trust between human warfighters and AI?* Transitioning from an AI-piloted *simulated* aircraft in a controlled virtual environment to an "AI" piloting a real \$50-million jet in a physical combat environment shrouded in the "fog of war" still entails an enormous technological leap (see the section on "**AI's Fundamental Gaps, Challenges, and Limitations**").

Perhaps the most intriguing questions of all are those of the *interpretability* and *understandability* of an AI's actions. While it may be "easy" to both interpret and understand Heron's winning AI's tactics against the human pilot in a controlled "proof of concept" trial ("run nose-to-nose and fire at close range"), we cannot expect this generally to be the case in more realistic future scenarios. Recall how, in the second game of the Go match between the AI that defeated Go world-champion Lee Sedol, the AI made a move so surprising that Lee 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," gasped Lee.<sup>148</sup> Both AlphaGo and its successor, AlphaZero, demonstrate a propensity for making "not typically 'human' moves," the nature of which is difficult if not impossible to discern (in human terms).<sup>149</sup> To be sure, an aerial dogfight (using a simulator) is not Go—among other things, the search spaces are different and not

---

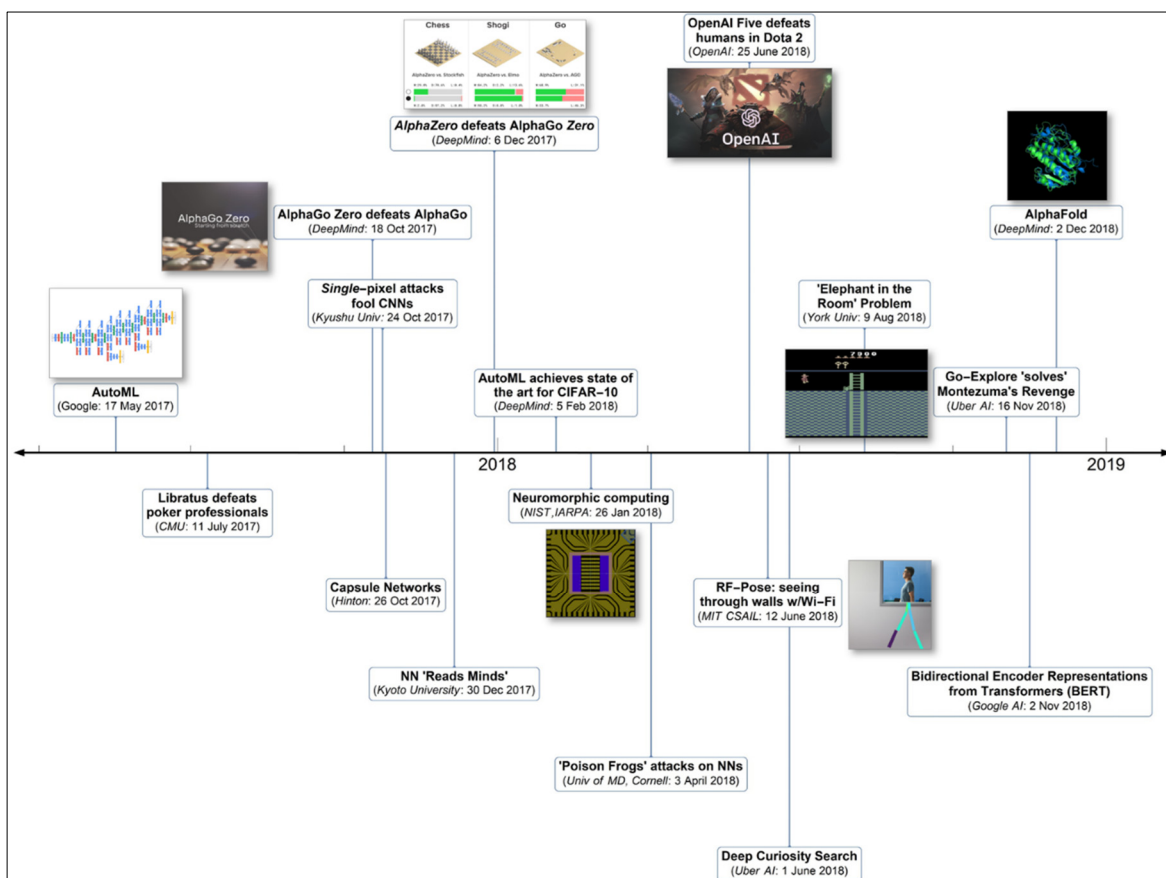
<sup>147</sup> Comments by Col. Daniel "Animal" Javorsek (program manager of DARPA's ACE program), 3:07:00 mark in video livestream of DARPA's ACE dogfight trials, 18-20 August 2020, <https://youtu.be/NzdhlA2S35w>.

<sup>148</sup> C. Metz, "The Sadness and Beauty of Watching Google's AI play Go," *Wired*, 11 March 2016.

<sup>149</sup> Matthew Sadler, *Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI*, New in Chess, 25 Jan. 2019.

commensurate in size (the “dogfight” space presumably trailing that of Go<sup>150</sup>)—but, in either case, there is the specter (perhaps even the *certainty*) of AIs surprising us with their “solution.” The deeper question is, how can we humans intelligently embrace and leverage AI’s inherent ability to discover novelty in unimaginably vast “possibility spaces”?<sup>151</sup>

Figure 14. A timeline of notable AI achievements, mid-year 2017 through end of 2018

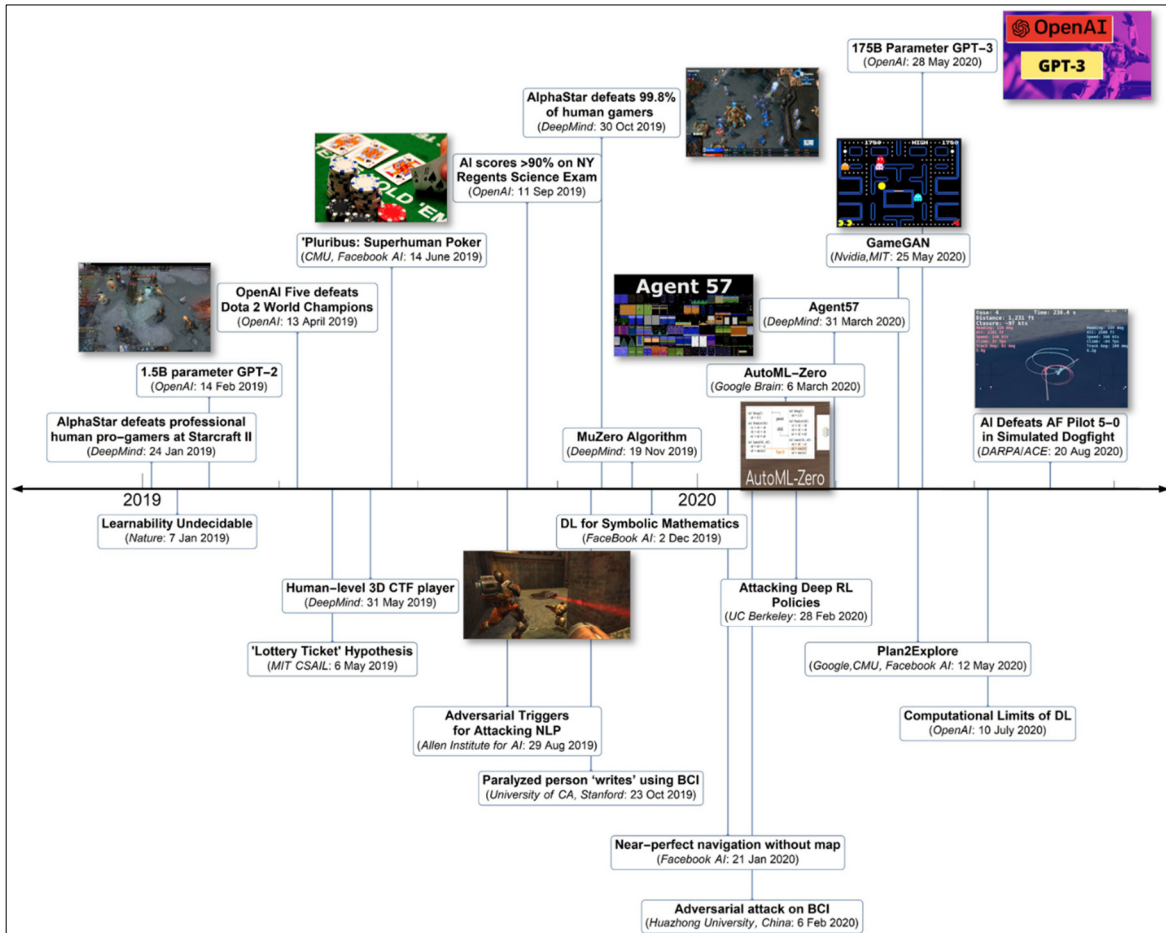


Source: CNA.

<sup>150</sup> Or does it? Given the vast number of state-space variables, sensor data about environmental conditions and (own and opponent) weapons load, flight and explosive yield characteristics information, and all of the flight controls that an AI “pilot” are in principle privy too, it is not hard to imagine a burst of micro-actions that result in a “kill” that are all but impossible for any human to follow and/or “understand.”

<sup>151</sup> Joel Lehman et al., “The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities,” 21 Nov. 2019, arXiv:1803.03453.

Figure 15. A timeline of notable AI achievements, 2019 through mid-year 2020



Source: CNA.

## AI “misses” (2017–2020)

While it is easy to be seduced into believing that everything AI “touches” turns to gold—particularly if one’s primary source of information consists of typically “hyped” news headlines and other secondhand accounts of bona fide research (see the “**Ever-present AI ‘Hype’ During 2017–2020**” section)—which is, of course, far from the truth. As with any other scientific endeavor, fundamental AI breakthroughs are impossible to predict, assessments of long-term impacts of prima facie breakthroughs may be overly optimistic (and/or subject to revised interpretation once the dust settles on preliminary announcements and follow-up research reveals problems remaining to be solved), and methods or technologies that conventional wisdom says are already mature simply fail when put to the real test. Leaving a more technical exposition of AI/ML’s research-level challenges to a later section (see “**AI’s**

**Fundamental Gaps, Challenges, and Limitations”**), we highlight here a few of AI’s “misses” during 2017–2020:

1. Amazon’s Rekognition facial recognition platform identifies 28 lawmakers as crime suspects.<sup>152</sup>
2. AI learns to associate colon cancer patients with specific clinics to which they were sent rather than the actual cancer (i.e., bias in electronic medical records).<sup>153</sup>
3. IBM Watson reportedly recommends “unsafe and incorrect” cancer treatments, and (in one review of recommendations for 656 colon cancer patients) matched those of experts only half the time.<sup>154</sup>
4. First pedestrian death with self-driving Uber vehicle in Tempe, Arizona.<sup>155</sup>
5. Two crashes and three deaths in one month by Tesla cars on “Autopilot.”<sup>156</sup>
6. Neural networks are easily fooled by strange poses of familiar objects.<sup>157</sup>
7. Study finds that the top seven (of 24 tested) forecasting methods are all basic statistical algorithms, not ML methods.<sup>158</sup>
8. Study finds no evidence of superior performance of ML over logistic regression for clinical prediction modeling.<sup>159</sup>
9. Study finds that 85 percent of neural-machine-translation AI projects ultimately fail.<sup>160</sup>
10. UK police’s facial recognition system found to have an 81 percent error rate.<sup>161</sup>

---

<sup>152</sup> Natasha Singer, “Amazon’s Facial Recognition Wrongly Identifies 28 Lawmakers, ACLU Says,” *New York Times*, 26 July 2018, <https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html>.

<sup>153</sup> Kayt Sukel, “Artificial intelligence: Looking beyond the hype,” *Medical Economics*, 27 Feb. 2019.

<sup>154</sup> Won-Suk Lee et al., “Assessing Concordance With Watson for Oncology, a Cognitive Computing Decision Support System for Colon Cancer Treatment in Korea,” *JCO Clinical Cancer Informatics* 2 (2018).

<sup>155</sup> D. Wakabayashi, “Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam,” *New York Times*, 19 March 2018, <https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html>.

<sup>156</sup> Tom Krisher, “3 crashes, 3 deaths raise questions about Tesla’s Autopilot,” *ABC News*, 3 Jan. 2020.

<sup>157</sup> Michael Alcorn et al., “Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects,” poster at the 2019 Conference on Computer Vision and Pattern Recognition, arXiv:1811.11553 [cs.CV].

<sup>158</sup> Spyros Makridakis et al., “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” *PLOS One*, 27 March 2018, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889>.

<sup>159</sup> E. Christodoulou et al., “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *Journal Clinical Epidemiology* 110 (2019).

<sup>160</sup> “Artificial Intelligence: Localization Winners, Losers, Heroes, Spectators, and You,” *Pactera*, 20 June 2019.

<sup>161</sup> Charlotte Jee, “London police’s face recognition system gets it wrong 81% of the time,” *MIT Technology Review*, 4 July 2020, <https://www.technologyreview.com/2019/07/04/134296/london-polices-face-recognition-system-gets-it-wrong-81-of-the-time/>.



## New AI “challenges” (2017–2020)

As a complement to the AI “misses” during 2017–2020 listed in the last section, we offer the following list of new AI *challenges* during this recent time period. Where “misses” are mostly specific and self-contained (i.e., “AI method X failed to accomplish task Y”), the examples below are broader and represent genuine challenges to accepted state-of-the-art and the ability of the AI/ML research community to sustain its heretofore increasingly rapid pace of development:

1. “Single-pixel,” “adversarial patch,” “elephant in the room,” “poison frog,” and “energy latency” adversarial attacks are able to fool convolutional neural networks.<sup>162</sup>
2. Reliability of current measures of progress in ML questioned.<sup>163</sup>
3. A state-of-the-art ML-translation system performs “better and more efficiently” when complexity is stripped.<sup>164</sup>
4. Mathematical proofs reveal fundamental limits of “learnability” and “knowability.”<sup>165</sup>
5. Meta-analysis study of ML methods finds that spurious samples cannot be eliminated without sacrificing a model’s ability to generate some data one actually wants to model.<sup>166</sup>
6. Multi-agent RL prone to irreproducibility, irreplicability, and increasing reliance on computing.<sup>167</sup>
7. Medical deep learning systems susceptible to adversarial attacks.<sup>168</sup>
8. National Institute of Standards and Technology (NIST) releases scathing review of inherent bias in 189 facial recognition algorithms.<sup>169</sup>
9. Review of ML research finds “troubling trends” in practices and scholarship.<sup>170</sup>

---

<sup>162</sup> *Single-pixel*: J. Su et al., “One pixel attack for fooling deep neural networks,” 24 Oct. 2017, arXiv:1710.08864; *adversarial-patch*: T. Brown et al., “Adversarial patch,” 17 May 2018, arXiv:1712.09665v2; *elephant in the room*: A. Rosenfeld et al., “The Elephant in the Room,” 9 Aug 2018, arXiv:1808.03305; *poison frog*: Ali Shafahi et al., “Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks,” arXiv:1804.00792v1; *energy latency*: Illia Shumailov et al., “Sponge Examples: Energy-Latency Attacks on Neural Networks,” 5 June 2020, arXiv:2006.03463.

<sup>163</sup> B. Recht et al., “Do CIFAR-10 Classifiers Generalize to CIFAR-10?” 1 June 2018, arXiv:1806.00451.

<sup>164</sup> M. Hutson, “AI researchers allege that machine learning is alchemy,” *Science*, 3 May 2018, <https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy>.

<sup>165</sup> David Wolpert, “Constraints on physical reality arising from a formalization of knowledge,” arXiv:1711.03499.

<sup>166</sup> B. Kegl et al., “Spurious samples in deep generative models: bug or feature?” 3 Oct. 2018, arXiv:1810.01876.

<sup>167</sup> P. H. Leal et al., “A Survey and Critique of Multi-agent Deep Reinforcement Learning,” arXiv:1810.05587.

<sup>168</sup> S. Finlayson et al., “Adversarial Attacks Against Medical Deep Learning Systems,” 4 Feb 2019, arXiv:1804.05296.

<sup>169</sup> P. Grother et al., *Face Recognition Vendor Test (FRVT)*, NIST, NISTIR-8280, Dec. 2019.

<sup>170</sup> Zachary C. Lipton and J. Steinhardt, “Troubling Trends in Machine Learning Scholarship,” paper presented at 2018 ICML, *The Debates*, arXiv:1807.03341.

10. Study finds that “the full driving task is too complex an activity to be fully formalized as a sensing-acting robotics system that can be explicitly solved through model-based and learning-based approaches in order to achieve full unconstrained vehicle autonomy.”<sup>171</sup>

## Growing “pushback” against AI/ML

As applications of AI/ML research become ever-more deeply and more frequently infused into social, cultural, political, and military domains, it is inevitable that the pushback against this happening will itself intensify, particularly when either the potential for or perceived existence of ethics violations are in play. Some “pushback” events that happened between the start of 2017 and August 2020 include (in chronological order):<sup>172</sup>

1. An open letter signed by 116 founders of robotics and AI companies from 26 countries urges the United Nations to “urgently address the challenge of lethal autonomous weapons (often called ‘killer robots’) and ban their use internationally” (notable signatories include Elon Musk, founder of Tesla, SpaceX, Neuralink, and OpenAI; Yoshua Bengio, leading deep-learning expert and founder of Element AI; and Jürgen Schmidhuber, leading deep learning expert and founder of Nnaisense).<sup>173</sup>
2. At the beginning of April 2018, Google employees urge the company’s CEO to pull out of Project Maven.<sup>174</sup> Two months later, Google announces that it will not renew its contract for Project Maven.<sup>175</sup>
3. AI researchers boycott a South Korean university’s research institute over its work on “killer robots.”<sup>176</sup>

---

<sup>171</sup> L. Fridman et al., “MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction with Automation,” *IEEE Access* 7, 1 July 2019.

<sup>172</sup> A more complete list than appears here (albeit one that includes events only through Oct. 2019) has been compiled by the AI Now Institute: <https://medium.com/@AINowInstitute/ai-in-2019-a-year-in-review-c1eba5107127>.

<sup>173</sup> “Killer robots: World’s top AI and robotics companies urge United Nations to ban lethal autonomous weapons,” Open Letter, Future of Life Institute, 20 Aug. 2017.

<sup>174</sup> S. Shane and D. Wakabayashi, “‘The Business of War’: Google Employees Protest Work for the Pentagon,” *New York Times*, 4 April 2018.

<sup>175</sup> Kate Conger, “Google Plans Not to Renew Its Contract for Project Maven, a Controversial Pentagon Drone AI Imaging Program,” *Gizmodo*, 1 June 2018.

<sup>176</sup> J. Vincent, “Leading AI researchers threaten Korean university with boycott over its work on ‘killer robots,’” *The Verge*, 4 April 2018.

4. Microsoft employees post open letter to prevent the company from bidding on the Joint Enterprise Defense Infrastructure (JEDI) contract, a \$10 billion project to build cloud services for DOD.<sup>177</sup>
5. Microsoft workers demand cancellation of a \$479 million contract with the Department of the Army for its Integrated Visual Augmentation System, which would apply Microsoft's HoloLens augmented reality technology to weapons development.<sup>178</sup>
6. San Francisco bans facial recognition software.<sup>179</sup>
7. Congressional legislation proposes to ban facial recognition from public housing (this marks the first time federal legislation has addressed limits on technology and tenants).<sup>180</sup>
8. The "Facial Recognition and Biometric Technology Moratorium Act" proposes to ban the use of facial recognition technology by federal law enforcement agencies.<sup>181</sup>
9. IBM, Amazon, and Microsoft all drop facial recognition research.<sup>182</sup>
10. On 30 June 2020, the Association for Computing Machinery (ACM)—the world's largest educational and scientific computing society—calls for an "immediate suspension of the current and future private and governmental use" of facial recognition technologies for "both technical and ethical reasons."<sup>183</sup>

## Unceasing AI "hype" during 2017–2020

Beginning with the *New York Times*' audacious proclamation in 1958 that the Navy had developed an "embryo" of a "thinking machine" (Rosenblatt's "perceptron") "**that it expects will be able to walk, talk, see, write, reproduce itself, and be conscious of its existence**,"<sup>184</sup> the AI hype engine has seldom seen a down day since. Headline highlights from 2017–2020 (in chronological order) include:

<sup>177</sup> Employees of Microsoft, "An Open Letter to Microsoft: Don't Bid on the US Military's Project JEDI," Medium, 12 Oct. 2018, <https://medium.com/s/story/an-open-letter-to-microsoft-dont-bid-on-the-us-military-s-project-jedi-7279338b7132>.

<sup>178</sup> "HoloLens for Good, Not War," Microsoft Workers 4 Good (@MSWorkers4), Twitter, 22 Feb. 2019 (5:00 pm), <https://twitter.com/MsWorkers4/status/1099066343523930112/photo/1>.

<sup>179</sup> Kate Conger et al., "San Francisco Bans Facial Recognition Technology," *New York Times*, 14 May 2019.

<sup>180</sup> Alfred Ng, "Facial recognition may be banned from public housing thanks to proposed law," CNet, 22 July 2019.

<sup>181</sup> C. Jee, "A new US bill would ban the police use of facial recognition," MIT Technology Review, 26 June 2020.

<sup>182</sup> "BM CEO's Letter to Congress on Racial Justice Reform," 8 June 2020, <https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/>.

<sup>183</sup> "Statement on Principles and Prerequisites for the Development, Evaluation and Use of Unbiased Facial Recognition Technologies," US Technology Policy Committee, ACM, 30 June 2020.

<https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf>.

<sup>184</sup> "New Navy Device Learns by Doing," *New York Times*, 8 July

1958, <https://timesmachine.nytimes.com/timesmachine/1958/07/08/83417341.html?pageNumber=25>.

1. “Facebook put cork in chatbots that created a secret language”<sup>185</sup>
2. “Computers are better than humans at reading”<sup>186</sup>
3. “Pretty sure Google's new talking AI just beat the Turing test”<sup>187</sup>
4. “Maryland researchers say they discovered ‘Holy Grail’ of machine learning”<sup>188</sup>
5. “Scientists Have Invented a Software That Can ‘See’ Several Minutes Into The Future”<sup>189</sup>
6. “An AI System Passed an Eighth-Grade Science Test. Can You?”<sup>190</sup>
7. “AlphaZero now showing human-like intuition in historical ‘turning point’ for AI”<sup>191</sup>
8. “This clever AI hid data from its creators to cheat at its appointed task”<sup>192</sup>
9. “A step closer to self-aware machines: ... engineers create robot that can imagine itself”<sup>193</sup>
10. “When bots teach themselves to cheat”<sup>194</sup>
11. “How to save America with artificial intelligence.”<sup>195</sup>

## COVID-19 and AI: *lessons learned for DOD?*

This short “bonus” section summarizes the response of the AI/ML research community to the ongoing (as of this writing, September 2020) coronavirus disease (COVID)–19 pandemic.<sup>196</sup> While the inclusion of this material may, on the face of it, seem odd and off-topic, the sudden appearance of a global threat (albeit a biological one, not military) for which a technology seemingly tailor-made to deal with its enormous complexities *already exists*—namely, AI—makes it a veritable laboratory in which to study AI’s applicability to real-world

---

<sup>185</sup> 31 July 2017: <https://www.cnet.com/news/what-happens-when-ai-bots-invent-their-own-language/>.

<sup>186</sup> 15 Jan 2018: <https://www.channelnewsasia.com/news/technology/alibaba-s-ai-software-surpasses-humans-in-reading-test-9863010>.

<sup>187</sup> 8 May 2018: <https://www.engadget.com/2018-05-08-pretty-sure-googles-new-talking-ai-just-beat-the-turing-test.html>.

<sup>188</sup> 29 May 2018: <https://www.washingtontimes.com/news/2018/may/29/maryland-researchers-say-they-discovered-holy-grail/>.

<sup>189</sup> 14 June 2018: <https://www.sciencealert.com/neural-network-software-predicting-human-actions-future-minutes>.

<sup>190</sup> 4 Sep 2019: <https://twnews.us/us-news/an-a-i-system-passed-an-eighth-grade-science-test-can-you>.

<sup>191</sup> 6 Dec 2018: <https://www.telegraph.co.uk/science/2018/12/06/deepminds-alphazero-now-showing-human-like-intuition-creativity/>.

<sup>192</sup> 31 Dec 2018: <https://techcrunch.com/2018/12/31/this-clever-ai-hid-data-from-its-creators-to-cheat-at-its-appointed-task/>.

<sup>193</sup> 30 Jan 2019: [https://www.eurekalert.org/pub\\_releases/2019-01/cuso-asc012819.php](https://www.eurekalert.org/pub_releases/2019-01/cuso-asc012819.php).

<sup>194</sup> 1 Aug 2019: <https://www.thepassivevoice.com/when-bots-teach-themselves-to-cheat/>.

<sup>195</sup> The Hill, February 2020, p. 416/v3.

<sup>196</sup> John Hopkins University School of Medicine, Coronavirus Resource Center, <https://coronavirus.jhu.edu/>.

operational problems. In turn, the concomitant successes and failures of these applications provide important “lessons learned” for DOD, as it continues to think its way through how best to invest in military applications of AI. The discussion in this section is necessarily brief and only touches on pertinent highlights. **Appendix D** contains a complete mindmap of all COVID-19 and AI-related stories covered on CNA’s “AI with AI” podcast between March and July 2020 (and includes embedded references).

Ever since the highly infectious and debilitating SARS-CoV-2 strain of coronavirus was first reported on 31 December 2019, in Wuhan, China,<sup>197</sup> the international response of the scientific community in general, and medical research community in particular, has been overwhelming.<sup>198</sup> Adding to this response was the “AI Challenge” issued by the White House in March to use AI (or other information technology innovations) to help understand and mitigate COVID-19’s impact on individuals, supply chains, and the economy.<sup>199</sup> As part of that challenge, the COVID-19 Open Research Dataset (CORD-19) was created to serve as a free repository of scholarly articles for the global research community.<sup>200</sup> COVID-19 literature has grown in much the same way as the pandemic itself—*exponentially*. For example, at CORD-19’s conception, it contained roughly 29,000 papers; as of August 2020, it houses more than 200,000 scholarly articles (including more than 100,000 with full text) about SARS-CoV-2 and related coronaviruses.

Other text-based resources include the National Institute of Health’s (NIH) COVID-19 Portfolio, which (as of July 2020) had 53,000-plus publications,<sup>201</sup> and Lawrence Berkeley National Laboratory’s COVID Scholar tracker, which (as of August 2020) contained in excess of 75,000 research papers.<sup>202</sup>

Jump-starting the participation of the AI and ML communities in COVID-19 related research were several major “AI challenges,” including one based on the COVID-19 Open Research Dataset (and hosted by Kaggle)<sup>203</sup> and another organized by the Roche Data Science Coalition

---

<sup>197</sup> Jingchun Fan et al., “Epidemiology of Coronavirus Disease in Gansu Province, China, 2020,” *Emerging Infectious Diseases* 26, no. 6 (June 2020), [https://wwwnc.cdc.gov/eid/article/26/6/20-0251\\_article](https://wwwnc.cdc.gov/eid/article/26/6/20-0251_article).

<sup>198</sup> Centers for Disease Control and Prevention, COVID-19 Research Articles Downloadable Database, <https://www.cdc.gov/library/researchguides/2019novelcoronavirus/researcharticles.html>.

<sup>199</sup> “Call to Action to the Tech Community on New Machine Readable COVID-19 Dataset,” Office of Science and Technology, 16 March 2020, <https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/>.

<sup>200</sup> The COVID-19 Open Research Dataset was created by the Allen Institute for AI in partnership with Georgetown University’s Center for Security and Emerging Technology (CSET), the Chan Zuckerberg Initiative, Microsoft Research, and the National Institutes of Health: <https://www.semanticscholar.org/cord19>.

<sup>201</sup> COVID-19 Portfolio, National Institute of Health, <https://icite.od.nih.gov/covid19/search/>.

<sup>202</sup> COVID Scholar, Lawrence Berkeley National Laboratory, <https://covid scholar.org/>.

<sup>203</sup> “COVID-19 Open Research Dataset Challenge: An AI challenge with AI2, CZI, MSR, Georgetown, NIH, and the White House,” <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>.

(RDSC).<sup>204</sup> The latter challenge—called the “UNCOVER COVID-19 Challenge”—presents a curated collection of datasets from 20 global sources and asks researchers to model solutions to key questions that were developed and evaluated by a global frontline of health care providers, hospitals, suppliers, and policy-makers.<sup>205</sup> This dataset (2 GB in size) is composed of a curated collection of more than 200 publicly available COVID-19–related datasets from sources like Johns Hopkins University School of Medicine, the World Health Organization (WHO), the World Bank, the *New York Times*, and many others. Overall, well over 2,000 AI/ML algorithms (for a variety of testbed tasks and problems) have been submitted for evaluation. These are but a few AI-related COVID-19 “challenges,” but of course do not include the hundreds, if not *thousands*, of one-off research efforts that COVID-19 has spawned.

Below, we summarize the conclusions of several recent major surveys of the efficacy of AI applications to COVID-19 (bold text is used to emphasize major takeaways):

1. Chen et al.<sup>206</sup> use databases such as Nature, Elsevier, Google Scholar, ArXiv, bioRxiv, and medRxiv to investigate the main scope and contributions of AI in combating COVID-19 from the aspects of disease detection and diagnosis, virology and pathogenesis, drug and vaccine development, and epidemic and transmission prediction. Linking various problems to AI technologies (which include some very specific connections—e.g., applying long short-term memory [see below], logistic regression, and “random forests” to vaccine development), the findings reveal that AI methods have been proposed for more than 18 broad categories of COVID-19–related problems, the most highly represented of which are image inspection, outbreak and transmission prediction, and proteomics, respectively. The survey also identifies four challenges faced by a majority of AI research efforts, including (1) **lack of available large-scale training data**; (2) **massively “noisy” data and the general propagation of misinformation and unverified rumors on social media sites**; (3) **a knowledge gap between AI experts and other computer scientists and medicine** (which is increasingly mitigated by initiatives and data repositories aimed at sharing data, models, and knowledge); and (4) **data privacy and human rights protection**.
2. A review published by BMJ of more than 14,000 papers posted at PubMed and Embase via Ovid, ArXiv, medRxiv, and bioRxiv through May 2020 revealed 107 studies describing 145 ML-based diagnosis and prognosis prediction models.<sup>207</sup> All models

---

<sup>204</sup> The Roche Data Science Coalition is a consortium organized by Hoffmann-La Roche Limited (Roche Canada) and includes Alberta Machine Intelligence Institute (Amii), doc.ai, NVIDIA, RGAX, Self-Care Catalysts, ThinkData Works Inc., and the Vector Institute: <https://www.rochecanada.com/>.

<sup>205</sup> United Network for COVID Data Exploration and Research (UNCOVER), COVID-19 Challenge, Kaggle, <https://www.kaggle.com/roche-data-science-coalition/uncover>.

<sup>206</sup> Jiango Chen et al., “A Survey on Applications of Artificial Intelligence in Fighting Against COVID-19,” 4 July 2020, <https://arxiv.org/pdf/2007.02202.pdf>.

<sup>207</sup> Laure Wynants et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, *BMJ* 369, <https://www.bmj.com/content/369/bmj.m1328>.

were rated at **high risk of bias**, mostly because of nonrepresentative selection of control patients, exclusion of patients who had not experienced the event of interest by the end of the study, **high risk of model overfitting**, and poor reporting (most reports did not include descriptions of the study population or intended use of the models, and calibration of the model predictions was rarely assessed). The review concludes, “**We do not recommend any of these reported prediction models for use in current practice.** Immediate sharing of well documented individual participant data from COVID-19 studies and collaboration are urgently needed to develop more rigorous prediction models and validate promising ones.... Methodological guidance should be followed because **unreliable predictions could cause more harm than benefit in guiding clinical decisions.**”

3. A review published in June 2020 in *Nature*— focusing mostly on natural language processing (NLP) tools to help biomedical researchers and clinicians to find relevant COVID-19 papers—found that while numerous NLP-based toolkits have emerged (of varying utility), **most tools are nascent and still in need of development; their utility remains largely unproven.**<sup>208</sup> Moreover, regarding the research papers themselves, **only about two-thirds of the papers in NIH’s COVID-19 portfolio are peer-reviewed** (a testament to the difficulty of peer-reviewed journals keeping pace with the rapid growth of papers).
4. A survey published in the *Stanford Social Innovation Review* in June 2020 highlights major data-related challenges (including data gaps and bias) and various ethics-related issues.<sup>209</sup> For example, the survey finds that much of the data about COVID-19 that the US Center for Disease Control and Prevention (CDC) and others are collecting and tracking are **incomplete and biased**; data on risk and mortality are not sufficiently disaggregated by sex, race, or ethnicity; data for racial and ethnic groups is incomplete, and terms and labels are inconsistent; and COVID-19 data tracking systems are not capturing data on immigrants and other marginalized populations. The survey emphasizes that algorithms that do not account for existing inequities risk making inaccurate predictions or worse, making the further point that although developers aim to make algorithms race-blind by excluding race as a metric, this can ignore or hide—rather than prevent—discrimination (e.g., algorithms that inform clinical decisions may use proxies such as preexisting conditions).
5. The National Endowment for Science, Technology and the Arts (NESTA)<sup>210</sup> published a comprehensive survey of the levels, evolution, geography, knowledge base and quality of AI research in the COVID-19 mission field using a novel dataset taken from the open

---

<sup>208</sup> Mathew Hutson, “Artificial-intelligence tools aim to tame the coronavirus literature,” *Nature*, 9 June 2020.

<sup>209</sup> Genevieve Smith and Ishita Rustagi, “The Problem With COVID-19 Artificial Intelligence Solutions and How to Fix Them,” *Stanford Social Innovation Review*, 5 June 2020.

<sup>210</sup> NESTA was created in 1998 from the first publically supported endowment in the UK and became an independent charity in 2012.

preprint sites ArXiv, bioRxiv and medRxiv.<sup>211</sup> Many of this survey's findings are noteworthy: AI remains underrepresented in this area compared to its presence in research outside of COVID-19; more than a third of the research focuses on predictive analyses of patient data and particularly medical scans; and AI and non-AI researchers working on COVID-19 tend to draw on different bodies of knowledge—less epidemiology, more pattern recognition, while AI's share of citations to computer science is five times higher than outside, and its share of citations to medicine is a third lower. Yet the survey's three main conclusions are particularly relevant for our present discussion (**our emphasis highlighted in boldface and red**):<sup>212</sup>

- **“The persistent underrepresentation of AI research in the COVID-19 mission field suggests some limitations in the generalizability of state-of-the-art algorithms into a new domain where data is fragmented, unreliable, and sensitive; mistakes could cost lives and explainability is at a premium.** Given this, and perhaps unsurprisingly, AI researchers have focused their efforts on computer vision and biomedical applications closer to their comfort zone at the risk of neglecting other important domains.”
- **“We find some evidence of silos between AI researchers and those in medical and biological science disciplines in tackling the pandemic.** AI researchers (and computer scientists) more broadly are sometimes accused of ‘solutionism,’ looking for technological fixes for complex societal problems such as predicting and controlling the spread of a pandemic, ensuring the sustainability of public health systems, or protecting the mental health of locked-down populations. **It will be difficult for them to develop truly effective technologies to tackle these challenges without tapping on the knowledge of other disciplines,** something that our analysis suggests is as common as it may be desired.”
- Citing a finding (appearing in an earlier portion of the survey) that AI papers tackling COVID-19 tend to receive fewer citations than other papers in the same topic—more precisely, the population of AI researchers active in COVID-19 research has a less established track record, proxied through the citations they have received in recent years—and referencing the issue of research quality, scholarship, and general reproducibility (or lack thereof) in the general AI research community, the survey finds that **“the comparatively low levels of citations received by AI researchers in our corpus, the weaker track record (in general) of AI researchers tackling COVID-19, the presence of a large number of research groups from unidentified institutions, and the large thematic jumps from some researchers into COVID-19 field suggest that similar risks may be present in AI research oriented toward COVID-**

---

<sup>211</sup> Juan Mateos-Garcia, J. Klinger, and K. Stathoulopoulos, *Artificial Intelligence and the Fight Against COVID-19*, NESTA, 15 June 2020, <https://www.nesta.org.uk/report/artificial-intelligence-and-fight-against-covid-19/full/>.

<sup>212</sup> *Ibid.*, p. 37.



19. Researchers, policymakers and practitioners need to develop strategies to validate contributions from new entrants into the COVID-19 mission field, while ensuring that new voices and ideas can still be heard.”

There is thus “the risk that researchers facing low barriers to entry into the field may produce low-quality contributions, making it harder to find valuable studies and discourage interdisciplinary contributions that could take longer to develop.”<sup>213</sup>

6. A systematic review of Embase via Ovid, and Medline via PubMed, bioRxiv, medRxiv, and ArXiv for published papers and preprints uploaded from 1 January 2020 to 24 June 2020, identified a total 952 studies that involved ML for COVID-19 detection and diagnosis using chest X-rays (CXR) and CT scans, of which 168 were included in the review after initial screening; **only 29 were found to be reproducible.**<sup>214</sup> Twenty of the studies are deep learning–focused, seven are focused on other ML approaches, and two incorporate both deep and non-deep ML methods. All studies reviewed have a **high or unclear risk of bias**; inappropriate control groups for non–COVID-19 patients; **use of small datasets**; unacknowledged use of “Frankenstein datasets” (i.e., datasets assembled from other datasets and redistributed under a new name); and no validation on external datasets (many papers give little attention to establishing the original source of the images). The review finds that **all papers suffer from methodological flaws and biases and report extremely optimistic results** (not warranted by content). “[We conclude] that *none of the developed models discussed are of potential clinical value.*”

The key takeaways from these surveys—small, incomplete, and/or biased datasets; a knowledge gap between AI researchers and medical experts; and unjustifiably optimistic reported results based on nascent technologies laden with methodological flaws—collectively serve as a microcosm of the fundamental issues facing DOD as it confronts the challenges of operationalizing AI.

COVID-19 demonstrates, in textbook fashion, that substantively applying AI to a problem that lies *outside the scope and expertise of typical ML researchers is hard, very hard*—particularly when relevant datasets are unavailable (or in short supply, not sufficiently representative of the requisite “training” space, and/or biased)—and succeeds best when researchers and problem-domain experts (in COVID-19’s case, medical practitioners, and in DOD’s case, military leaders, policy-makers, and warfighters) are all aligned by common goals, priorities, and methodology and, in DOD’s case, are also aligned by a *common language* (as will be argued in this paper’s final section).

We conclude this section by mentioning two additional, *implicit* “lessons learned” from the AI/ML research community’s response to the COVID-19 pandemic:

---

<sup>213</sup> Ibid., p. 3.

<sup>214</sup> Michael Roberts et al., “Machine learning for COVID-19 detection and prognostication using chest radiographs and CT scans: a systematic methodological review,” 1 Sept. 2020, <https://arxiv.org/abs/2008.06388>.

1. **The demonstrable utility of “low-hanging fruit.”** The most immediately “useful” applications have, for the most part, been based not on cutting-edge methods, but on “yesteryear’s” tried-and-true approaches. For example, one of the best performing algorithms thus far developed (as of August 2020) predicts new state-wide outbreaks of COVID-19 using a combination of solid data and an “old-fashioned” Bayesian inference model (developed years ago).<sup>215</sup> The researchers also underscore the fact that while deliberating on what method to use to process real-time data streams (their method requires combining Google searches with many other kinds of data), they settled not on a SOTA ML-based semantic clustering method but, rather, a “simple” keyword-based filtering scheme.<sup>216</sup> On a broader playing field, one of the most popular ML-based tools used by both AI and medical researchers is a (still-growing) set of “vanilla” (i.e., fairly conventional applications) NLP-derived *meta*-tools designed to help researchers keep abreast of the exponentially growing body of technical literature on COVID-19.<sup>217</sup>
2. **Application is not synonymous with operationalization.** The surveys referenced above only touch on the basic challenges associated with operationalizing seemingly tailor-made applications. To further emphasize this lesson, and as a prelude to the ensuing discussion in the following section, we highlight two recent (non-COVID-related) examples: (1) GoogleAI’s ostensibly laudable ML-based system for breast cancer mammogram screening (published in *Nature* in January 2020)<sup>218</sup>—that outperforms human readers—was criticized by medical practitioners for effectively solving the wrong problem (“When you subject symptom-free people to mammograms ... you’ll end up finding a lot of things that look like cancer but will never threaten anyone’s life”);<sup>219</sup> and (2) a study by Google Health—the first to look at the impact of a deep-learning tool in real clinical settings—reveals that even the most accurate AIs can actually make things worse if not tailored to the clinical environments in which they will work.<sup>220</sup>

---

<sup>215</sup> Nicole E. Kogan et al., “An Early Warning Approach to Monitor COVID-19 Activity with Multiple Digital Traces in Near Real-Time,” 3 July 2020, <https://arxiv.org/pdf/2007.00756.pdf>.

<sup>216</sup> “While a machine learning-based semantic clustering method like Guided Latent Dirichlet Allocation (LDA) may deliver more comprehensive results (e.g., through identifying co-occurring and unknown terms), controlling the ratio between false positives and false negatives requires extensive experimental work and expert knowledge,” *ibid.*, p. 14

<sup>217</sup> See (1) Amazon Web Services’ COVID-19-Search (<https://arxiv.org/pdf/2007.12731.pdf>); (2) Aminer.org’s COVID-19 Knowledge Graph (<https://covid-19.aminer.cn/>); (3) Allen Institute for AI’s COVID-19 Claim Verification (<https://arxiv.org/pdf/2004.14974.pdf>); (4) Lawrence Berkeley National Laboratory’s COVIDScholar (<https://covid scholar.org/>); and (5) Johns Hopkins University’s 2019 Novel Coronavirus Research Compendium (<https://primer.ai/blog/science-vs-covid-19/>) among many others (see <https://www.cna.org/CAAI/audio-video>).

<sup>218</sup> Scott McKinney et al., “International evaluation of an AI system for breast cancer screening,” *Nature* 577 (1 Jan. 2020), <https://www.nature.com/articles/s41586-019-1799-6>.

<sup>219</sup> Christie Aschwanden, “Artificial Intelligence Makes Bad Medicine Even Worse,” *Wired*, 10 Jan 2020.

<sup>220</sup> Emma Beede et al., “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy,” *Conference on Human Factors in Computing Systems*, April 2020.

## Specter of a *stall-in-progress*?

An earlier section (see “**Short history**”) recounts how AI research has undergone at least two “winters.” The first winter—labeled “Dark Period 1” in Figure 4—in the 1970s was precipitated by the publication of the book *Perceptrons*, which demonstrated that the basic design of the artificial neural-network design architecture then in vogue entailed significant computational limitations. The second winter—beneath the area labeled “Dark Period 2” in Figure 4—spanned roughly a decade, between the mid-1990s and mid-2000s, and was precipitated by the relative *slowness* of the computer processors and *limited memory storage* of circa 1990s-era computers and the absence of computationally inexpensive learning algorithms for NNs with many hidden layers (which were introduced only in 2006). But these were by no means the only winters. Another significant winter, at least one that took place within the Japanese S&T sector, was the massive failure of Japan’s circa 1990s so-called—and, at the time, widely heralded—“Fifth Generation Project” (5GenP), whose goal was to develop a “thinking machine” (it was not called AGI at the time, but that was essentially the idea).<sup>221</sup> After spending more than \$400 million on the ten-year project, during which time the technologies being developed for it were leapfrogged by advances made by the computer industry as a whole (and generally falling far short of its goal), the Japanese government pulled the plug on any further development. It later admitted that it was the concomitant training that its engineers received in advanced computer science, as a side-benefit of participating in the 5GenP, rather than achieving 5GenP’s goals, that are the real legacy of the 1990s effort.<sup>222</sup> A number of other “mini” winters have also occurred.<sup>223</sup>

*Is there evidence of an impending new AI winter (a looming “Dark Period 3” addition to an expanded version of Figure 4)?* Some researchers have recently gone on record suggesting that such a winter has already begun. For example, Francois Chollet, creator of Keras, posted this note on his Twitter page in 2018: “Today more people are working on deep learning than ever before, around two orders of magnitude more than in 2014. And the rate of progress as I see it is the slowest in five years. Time for something new.”<sup>224</sup> Gary Marcus, an AI researcher at New York University, has opined that, “By the end of the decade there was a growing realization that current techniques can only carry us so far.”<sup>225</sup> And AI researcher Filip Piekiewicz has

---

<sup>221</sup> Edward Feigenbaum and H. Shrobe, “The Japanese national Fifth Generation project: Introduction, survey, and evaluation,” *Future Generation Computer Systems* 9 (July 1993).

<sup>222</sup> Andrew Pollack, “‘Fifth Generation’ Became Japan’s Lost Generation,” *New York Times*, 5 June 1992.

<sup>223</sup> Nilsson, *The Quest for Artificial Intelligence*. See also, “AI Winter,” Wikipedia, [https://en.wikipedia.org/wiki/AI\\_winter](https://en.wikipedia.org/wiki/AI_winter).

<sup>224</sup> Chollet, Twitter, 10 Sep 2018.

<sup>225</sup> Quoted in Sam Sheard, “Researchers: Are we on the cusp of an ‘AI winter’?” BBC News, 12 Jan 2020.

suggested that, “Fears of an impending winter are hardly skin deep. Deep learning has slowed in recent years.”<sup>226</sup>

In an essay posted on his personal website, Piekniewski, after discussing several aspects of AI winter and its hype, concludes the following:

Predicting the AI winter is like predicting a stock market crash—impossible to tell precisely when it happens, but almost certain that it will at some point. Much like before a stock market crash, there are signs of the impending collapse, but the narrative is so strong that it is very easy to ignore them, even if they are in plain sight. In my opinion, there are such signs of a huge decline in deep learning (and probably in AI in general, as this term has been abused “ad nauseam” by corporate propaganda) already visible. Visible in plain sight, yet hidden from the majority by an increasingly intense narrative. How “deep” will this winter be? I have no idea. What will come next? I have no idea. But I’m fairly positive it is coming, perhaps sooner rather than later.<sup>227</sup>

But rather than pontificate on the likelihood and possible implications of another AI winter, we instead summarize growing evidence that suggests that, at the very least, AI and ML research may be seeing a *stall-in-progress*, wherein genuine new advances, in technique and/or performance, appear to be slowing. If true, this has serious implications for DOD vis-à-vis near-term and future strategies for applying and operationalizing AI technologies. There has also been a spate of papers questioning the empirical rigor inherent in, and generally declining state of, recent ML scholarship.<sup>228</sup>

A number of probing assessments of general AI research practices, along with reviews and surveys of SOTA methods (compared to “basic statistical” methods) and trade-offs between computational requirements (for training AI systems) and performance, have recently been published. Key takeaways listed in order of the respective reviews’ dates of publication include the following:

1. A comparison of the performances of 26 different forecasting methods using a large subset of more than 1,000 monthly time-series datasets used in the M3 Competition<sup>229</sup> **found that the seven most accurate methods are basic statistical methods, not**

---

<sup>226</sup> Quoted in Smriti Sriastava, “AI Winter is Coming: Hear from Experts, What Could Possibly Happen?” Analytics Insight, 13 Jan. 2020, <https://www.analyticsinsight.net/ai-winter-coming-hear-experts-possibly-happen/>.

<sup>227</sup> Filip Piekniewski, “AI Winter Is Well on Its Way,” Piekniewski’s blog: On limits of deep learning and where to go next with AI, 28 May 2018, <https://blog.piekniewski.info/2018/05/28/ai-winter-is-well-on-its-way/>.

<sup>228</sup> For example: (1) Zachary Lipton and Jacob Steinhardt, “Troubling Trends in Machine Learning Scholarship,” 26 July 2018, arXiv:1807.03341; (2) D. Sculley et al., “Winner’s Curse? On Pace, Progress, and Empirical Rigor,” ICLR Workshops, 2018, <https://openreview.net/pdf?id=rJWF0FywF>; and (3) Denny Britz, “AI Research, Replicability, and Incentives,” 17 June 2020, <https://dennybritz.com/blog/ai-replication-incentives.pdf>.

<sup>229</sup> M3 Competitions (i.e., Makridakis Competitions) are a series of open competitions organized by teams led by forecasting researcher Spyros Makridakis and intended to evaluate and compare the accuracy of different forecasting methods. Note that Makridakis is lead author of the following reference.

**ML.**<sup>230</sup> The review concludes that, “ML methods are not a panacea that would automatically improve forecasting accuracy. Their capabilities can easily generate implausible solutions, leading to exaggerated claims of their potentials and must be carefully investigated before any claims can be accepted.”

2. Google Brain’s large-scale study of Generative Adversarial Networks (GANs)—the class of ML algorithms that sits at the heart of today’s rapidly improving quality of image– and video–based “deep fakes”<sup>231</sup>—found **no evidence that any of today’s SOTA algorithms outperform the original GAN algorithm**, introduced in 2014.<sup>232</sup> Another review found that when an NN architecture called long short-term memory (LSTM), introduced in 1997, was properly trained, its performance **matched that of supposedly more advanced architectures developed two decades later.**<sup>233</sup>
3. A meta-analysis of information retrieval (IR) algorithms used in search engines that examined every IR-related publication from 2005 to 2018 concluded that the **“high-water mark ... was actually set in 2009, and no reported results since then (neural or otherwise) come close.”**<sup>234</sup> The review concludes with the assertion that, “while neural networks no doubt represent an exciting direction in information retrieval, we believe that at least some of the gains reported in the literature are illusory.”
4. A comprehensive review of all NN-based recommendation systems (similar to those used by media streaming services) that were published at prestigious scientific conferences between 2015 and 2018 concluded that **“11 out of the 12 reproducible neural approaches can be outperformed by conceptually simple methods, e.g., based on the nearest-neighbor heuristic or linear models.”**<sup>235</sup> Moreover, the review finds that none of the NN-based methods was consistently better than already-existing learning-based techniques; indeed, six failed to outperform much simpler, non-neural algorithms developed years before, when techniques were fine-tuned, suggesting that **“while many papers claiming to make advances over the state of the art were published, these mostly seemed to amount to phantom progress.”**

---

<sup>230</sup> S. Makridakis et al., “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” PLOS One, 27 March 2018, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889>.

<sup>231</sup> Given a training set, GANs learn to generate new data with the same statistics as the training set. They generally consist of two separate networks: one network generates candidates, while the other network evaluates them. GANs were introduced by Ian Goodfellow et al., “Generative Adversarial Networks,” 10 June 2014, <https://arxiv.org/abs/1406.2661>.

<sup>232</sup> Mario Lucic et al., “Are GANs Created Equal? A Large-Scale Study,” 29 Oct. 2018, arXiv:1711.10337v4.

<sup>233</sup> Mathew Hutson, “Core progress in AI has stalled in some fields,” *Science* 368 (29 May 2020).

<sup>234</sup> Wei Yang et al., “Critically Examining the ‘Neural Hype’: Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2019)*, July 2019, <https://arxiv.org/pdf/1904.09171.pdf>.

<sup>235</sup> Maurizio Dacrema et al., “A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research,” *ACM Transactions on Information Systems*, 27 Aug. 2020, arXiv:1911.07698v2.

5. As part of an analysis of methods to train image recognition models to be immune to "adversarial attacks," researchers at Carnegie Mellon University found that, contrary to intuition (which suggests that today's SOTA methods are the most efficient), **when subtly tweaked, the performance of all of today's methods are roughly on par with each, and with that of one of the earliest adversarial training methods.**<sup>236</sup>
6. A recent review of NN *pruning* (i.e., the task of reducing the size of a network by removing parameters in hopes of finding a design that achieves roughly the same performance as the original network but at a computationally smaller training cost) found that a side-by-side comparison of 81 different SOTA pruning algorithms showed **no clear evidence of any performance improvements over a ten-year period.**<sup>237</sup>
7. A survey of deep metric learning (whose goal is to map data to an embedding space in such a way that similar data are close together and dissimilar data are far apart) found that when a dozen SOTA methods are compared on equal footing on an image retrieval task, **accuracy has not improved since 2006.**<sup>238</sup>

All the surveys listed above report their results based on comparing the *relative performance* of SOTA methods to either older ML-based algorithms or basic statistical routines. What about measuring performance gains in absolute terms, or assessing trade-offs between accuracy and computational demands? The answers to these questions are not so straightforward, and may even appear to be in partial conflict, but they are no less worrisome in terms of likely near- to mid-term progress in SOTA ML methodology.

On the one hand, OpenAI recently published an analysis showing that the amount of computation needed to train a neural net to achieve the same level of performance on ImageNet classification as demonstrated by AlexNet in 2012 has been decreasing by a factor of two every 16 months.<sup>239</sup> In other words, compared to 2012, it now takes 44× less computation in terms of TeraFLOP-days (where a FLOP = a floating-point operation, and a TeraFLOP = one trillion FLOPs) to train a neural network to the 2012 level of AlexNet.<sup>240</sup> (By contrast, Moore's Law would yield an 11x cost improvement over this period.) The results suggest that for AI tasks with high levels of investment (in terms of researcher time and/or computational needs), algorithmic efficiency might outpace gains from hardware efficiency.

---

<sup>236</sup> Eric Wong, L. Rice, and J. Kolter, "Fast is Better than Free: Revisiting Adversarial Training," paper presented at the International Conference on Learning Representations (ICLR), Jan. 2020, arXiv:2001.03994v1.

<sup>237</sup> Davis Blalock et al., "What is the State of Neural Network Pruning?" *Proceedings of the 3rd MLSys Conference*, Austin, TX, March 2020, <https://arxiv.org/pdf/2003.03033.pdf>.

<sup>238</sup> K. Musgrave et al., "A Metric Learning Reality Check," 24 July 2020, <https://arxiv.org/pdf/2003.08505.pdf>.

<sup>239</sup> The ImageNet Challenge, an annual contest run by Stanford University, was famously won in 2012 by AlexNet, which was the *only* NN-based entry that year and the first NN-based algorithm to win the challenge.

<sup>240</sup> D. Hernandez and T. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," OpenAI, May 2020, [https://cdn.openai.com/papers/ai\\_and\\_efficiency.pdf](https://cdn.openai.com/papers/ai_and_efficiency.pdf).

On the other hand, a recent analysis of the trade-offs between accuracy and computational training demands finds that **deep learning has already entered an era of diminishing returns**.<sup>241</sup> Specifically, the researchers analyzed more than 1,000 papers from the ArXiv preprint server (see “**Recent Trends**”) as well as other benchmark sources to understand the connection between deep-learning performance and computation, focusing attention on five specific domains: *image classification, object detection, question answering, named entity recognition, and machine translation*.

The review’s researchers find that three years of algorithmic improvement is roughly equivalent to a 10x increase in computing power and that progress in *all five areas* “**show large increases in computational burdens with relatively small improvements in outcomes.**” More specifically (and looking toward the future), the researchers estimate that, with respect to image classification and under the most optimistic assumptions, **it will require an additional 105x more computing power to achieve an error rate of 5 percent for ImageNet** (the current error rate, as of August 2020, is 11.5 percent). Projecting the computational power needed to hit various max-error-rate benchmarks for each of the five problem domains, the review concludes that, “**Along current trends, it will not be possible for deep learning to hit these benchmarks.** Instead, fundamental re-architecting is needed to lower the computational intensity so that the scaling of these problems becomes less onerous.” The review concludes that “deep learning’s prodigious appetite for computing power imposes a limit on how far it can improve performance in its current form, particularly in an era when improvements in hardware performance are slowing,” and that continued progress “will require dramatically more computationally efficient methods, which will either have to come from changes to deep learning or from moving to other machine-learning methods.”<sup>242</sup>

## Possible implications for DOD

The two most important takeaways from this section and from the last (see “**COVID-19 & AI**”) are (1) *deep learning may already have entered (or is soon to enter) an era of diminishing returns*, in which ever-increasing computational resources and/or refinements in algorithmic technique yield relatively marginal gains in performance; and (2) that *applying AI to problems that lie outside the domains for which they were originally developed is very hard to do*, particularly when requisite datasets are either incomplete, in short supply, biased, or simply unavailable (e.g., how does one train an ML-based jamming system when the data describing an adversary’s radars’ wartime emissions modes may be unknown).<sup>243</sup> To which we must also tack on the list of fundamental challenges (both perennially persistent and newly appreciated in recent years, as discussed in the sections “**Fundamental Gaps, Challenges, and**

---

<sup>241</sup> Neil C. Thompson et al., “The Computational Limits of Deep Learning,” 10 July 2020, arXiv:2007.05558v1.

<sup>242</sup> Ibid., p. 15.

<sup>243</sup> Shixun You, M. Diao, and L. Gao, “Deep Reinforcement Learning for Target Searching in Cognitive Electronic Warfare,” *IEEE Access* 7 (18 March 2019), <https://ieeexplore.ieee.org/abstract/document/8668391>.

**Limitations”** and **“New AI Challenges: 2017–2020”**) associated with applying, developing, and deploying AI/ML systems.

For DOD, these takeaways and challenges collectively portend a looming transitional period during which its AI/ML-related S&T portfolio investment strategies will shift from leveraging “low-hanging fruit” applications (e.g., using PyTorch or TensorFlow to help intelligence analysts parse satellite imagery) to focusing more on operationalizing existing SOTA AI/ML methods (whose performance is already “good enough”). This requires a deeper commitment to understanding what “operationalizing AI” really means for the military, along with a greater need of adopting a common language that can be shared among AI/ML researchers and military leaders, policy-makers, and warfighters. The next, and last, section of this paper discusses one possible approach to developing such a language.



# Moving From “Understanding” to Operationalizing AI

---

## Theories of general *human* intelligence

“AI will now proceed along two parallel path: (1) specialized systems, and (2) habile systems ... general, intelligent systems ... having general skill.”

— Nils J. Nilsson, *Eye on the Prize* (1995)

Having begun this paper by asking, “What is artificial intelligence?” it is only fitting that we begin our concluding narrative by posing a deceptively simpler form of this question, “What is *intelligence*?” There are two heuristic motivations for asking this question.

First, if—as the main title of this final section suggests—the goal is to move away from “understanding AI” (to avoid the cacophonous disagreements that continued debate about how to best *define* AI engenders) and toward *operationalizing* its potential military applications, it behooves us to also move away from asking what AI *is*, toward what we require and/or expect AI to *do*—that is, to inquire about the kind of performance(s) that an engineered AI system can and/or should demonstrate.

Second, if “operationalizing AI” involves, as it has already started to, not just “single task” solutions (i.e., the “low-hanging fruit” class of applications discussed earlier in this paper), but also the “baby AGI”-like synergistic “AI-system”-of-“AI-system” clusters of tasks and behaviors (whose coupled performance will increasingly stretch across multiple problem domains and complex dynamic environments), we must also move away from using myopic, single-task performance measures (like image classification accuracy) and toward developing more holistic, *psychometric* evaluations of general abilities that probe multiple, simultaneous behavioral dimensions.<sup>244</sup> So, taking a cue from these two heuristics, our first task is to better understand what *intelligence* consists of, be it artificial, human, or some other heretofore undiscovered variant.

Of course, we do not have the space to give much more than a cursory look at a vast field of research. Our only goal here is to provide a sufficient context to appreciate why we have chosen to use the so-called CHC taxonomy of cognitive abilities (introduced in the next section) as the

---

<sup>244</sup> José Hernández-Orallo, *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*, New York: Cambridge University Press, 2017.

backbone of a “template of a framework” for operationalizing military applications, proposed in the final section of this paper.

Following Mackintosh,<sup>245</sup> we trace the first theories of intelligence to the earliest Greek philosophers—such as Plato, for whom intelligence was rooted in the love of learning, and the love of truth. Moving quickly through the centuries, from the 1600s–1700s, during which time psychology slowly emerged as a discipline separate from philosophy; to the 1800s, which saw the rise of several prominent schools of psychology throughout Europe and the US and the first “theories of intelligence” (e.g., by Wilhelm Wundt, Sigmund Freud, and William James); to the early to mid-1900s, which gave rise to the first intelligence tests and (then) new statistical techniques for administering and evaluating them (starting with Charles Spearman’s circa 1904 concept of general intelligence—wherein individuals are assumed to possess a certain general level of intellectual ability—which he denoted by a single italicized letter, *g*, and to which IQ tests still effectively assign a single numerical score) and inklings of a finer structure to Spearman’s *g* were first stirring; to the latter half of the 1900s, when full-blown “multiple intelligences” theories began to appear, and the study of “human intelligence” as a field started cross-fertilizing with, and leveraging, findings from the neuroscience community.

It is this latter approach to describing, and measuring, intelligence—namely, the concept of *multiple intelligences*—that is central to our narrative. Just as E. O. Wilson has argued that science should generally work toward achieving *consilience*—that is, toward integrating diverse domains of knowledge<sup>246</sup>—so too a number of recent theories of intelligence have advocated the idea that “intelligence” is best described as a unity of separate abilities. Early proponents of this idea included Louis Thurstone, who in the 1930s suggested that intelligent behavior does not arise from a general factor but instead emerges from different “primary mental abilities”; Philip Vernon, who in the 1960s proposed a hierarchical group factor theory of the structure of human intellectual abilities based on factor analysis (which effectively splits Spearman’s *g* into *major*, *minor*, and *specific-factor* tiers); and Joy Guilford, who in the 1950s introduced the Structure of the Intellect model, an early precursor to the CHC framework discussed in the following section.<sup>247</sup>

---

<sup>245</sup> Unless otherwise noted (and referenced), key ideas introduced in this section derive from N. J. Mackintosh, “History of Theories and Measurement of Intelligence” (pp. 3–19) in *The Cambridge Handbook of Intelligence*, ed. R. Sternberg and S. Kaufman, New York: Cambridge University Press, 2011.

<sup>246</sup> Edward O. Wilson, *Consilience: The Unity of Knowledge*, New York: Knopf, 1998.

<sup>247</sup> The SOI is a three dimensional model that consists of sets of *operations* (e.g., cognition, memory, and evaluation), *content* (e.g., figural, symbolic, and semantic), and *products* (e.g., units, classes, and relations), and yields a total of 120 possible intellectual factors. See J. P. Guilford, “The structure of intellect,” *Psychological Bulletin* 53 (1956).

More-recent approaches include the following:

- *Bloom's taxonomy*, which is a framework introduced in 1956 to classify educational learning using six key categories of *knowledge, comprehension, application, analysis, synthesis, and evaluation*.<sup>248</sup> This taxonomy was reconceptualized in the early 2000s to reflect a more dynamic categorization using verbs and gerunds to label categories and subcategories (instead of using nouns, as in Bloom's original taxonomy). These action words are used, in turn, to describe the cognitive processes by which "intelligence" works with knowledge (e.g., remembering, recognizing, summarizing, explaining, etc.). The revised taxonomy remains grounded in knowledge, but adds a separate class of knowledge used in cognition (e.g., *factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge*).<sup>249</sup>
- Howard Gardner's *theory of multiple intelligences*, which he introduced in his 1983 book, *Frames of Mind*.<sup>250</sup> His theory directly challenged the notion of a single, all-encompassing type of intelligence (i.e., Spearman's *g*), broadening it to consist of eight different categories of intelligence:<sup>251</sup>
  - *Verbal/Linguistic Intelligence*, which describes the facility to deal with spoken and written language, the ability to learn languages, and capacity to use language to accomplish certain goals.
  - *Logical-Mathematical Intelligence*, which refers to the capacity to think logically, perform mathematical calculations, and generally engage in abstract thought.
  - *Spatial Intelligence*, which includes the ability to recognize and mentally manipulate images and patterns in physical space.
  - *Bodily-Kinesthetic Intelligence*, which describes the ability to control one's body (whole and part), and the capacity to interact with the physical environment.
  - *Musical Intelligence*, which refers to abilities having to do with patterns of sounds, rhythms, and tones, including the capacity to compose and perform.
  - *Interpersonal Intelligence*, which is the capacity to understand the intentions, motivations, moods, feelings, temperaments, and desires of other people.

---

<sup>248</sup> Benjamin S. Bloom, *Taxonomy of Educational Objectives: Cognitive Domain*, Boston: Addison-Wesley, 1956.

<sup>249</sup> Lorin W. Anderson et al., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, New York: Pearson, 2000.

<sup>250</sup> Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences*, New York: Basic Books, 1983.

<sup>251</sup> Gardner's *Frames of Mind* contains the first seven on this list. The last two "intelligences" were added by Gardner in later years. See, for example, Howard Gardner, *Multiple Intelligences: New Horizons*, New York: Basic Books, 2008.

- *Intrapersonal Intelligence*, which describes the ability to understand oneself and the general capacity for introspection.
- *Naturalist Intelligence*, which includes the ability to recognize elements of, and make distinctions among, objects in the world and a general capacity to nurture and relate information to one's natural surroundings.
- *Existential Intelligence*, which Gardner later added (as a possibly “useful construct”) to account for the “spiritual” dimension of intelligence.
- Marvin Minsky’s *The Society of Mind* (TSoM) is a multidimensional agent-based theory of natural intelligence introduced in a 1988 book of the same title.<sup>252</sup> It is a bit ironic that Minsky—whose 1969 book, *Perceptrons* (co-authored with Seymour Papert) was instrumental in precipitating the first “dark period” of AI research (i.e., “**Neural networks and deep learning**”—should also appear in the closing pages of this paper as an illustrative example of how a theory of “multiple intelligences” can be used to develop a framework for operationalizing AI.

Briefly described, Minsky’s theory, presented over the course of 250-plus self-contained essays (and organized around 30 chapters, each of whose entries is focused on a single broad component of intelligence), is that human intelligence is a naturally emergent dynamic consequence of myriad entwined simultaneous interactions of simple parts (which he calls “agents”), which are themselves mindless. The web of interactions is what constitutes the “society of mind.” Agents include those that facilitate understanding, adjudicate parts and wholes, enable seeing and believing, provide the capacity to brainstorm, form and retrieve memories, deal with emotion, solve problems, and so forth.

## Cattell–Horn–Carroll (CHC) Framework

The Cattell-Horn-Carroll (CHC) theory of cognitive abilities is, to date, the most comprehensive psychometric-based model for understanding the structure of cognitive abilities and human intelligence.<sup>253</sup> Because of the extensive empirical support it has received in the research literature, it is widely used as the consensus foundation for developing, administering, and interpreting tests of intelligence and cognitive abilities. CHC’s origins date back to the early 1940s, when Raymond Cattell (who was a student of Spearman) and his (then) student John Horn proposed two types of intelligence, *Gf* and *Gc*, which represent, respectively, “fluid intelligence” and “crystallized intelligence:”<sup>254</sup> *Gf* refers to inductive and deductive reasoning

---

<sup>252</sup> Marvin Minsky, *The Society of Mind*, New York: Simon and Schuster, 1988. The contents of the entire book is (as of September 2020) freely available at <http://aurellem.org/society-of-mind/index.html>.

<sup>253</sup> Dawn P. Flanagan and Shauna G. Dixon, “The Cattell-Horn-Carroll Theory of Cognitive Abilities,” in *Encyclopedia of Special Education*, ed. Cecil R. Reynolds et al., Hoboken, NJ: Wiley, 2013.

<sup>254</sup> Raymond B. Cattell, *Personality and Motivation Structure and Measurement*, World Book, 1957.

abilities (that may be influenced by biological and neurological factors as well as learning achieved via interaction with environment) and includes abilities to “act quickly,” “think,” “solve problems,” and retrieve short-term memories; *Gc* refers to a class of abilities whose roots are in learning and acculturation (as reflected via general knowledge acquisition, verbal skills, and language skills). Cattell and Horn’s original formulation includes around 100 different abilities organized around these two main classes.

In the 1960s, John Horn added four additional classes of general abilities to the basic *Gf-Gc* model:<sup>255</sup> (1) visual perception and processing (*Gv*), short-term memory acquisition and retrieval (*Gsm*), long-term memory acquisition and retrieval (*Glr*), and cognitive processing speed (*Gc*).<sup>256</sup> In the 1990s, Horn also added “reaction time in making decisions” (*Gt*). With Horn’s contributions, along with a few others that eventually expanded the list to include two additional quantitative (*Gq*) and broad reading and writing abilities (*Grw*), the Cattell-Horn *Gf-Gc* theory matured and assumed its present form (circa 2000).

In 1993, John Carroll introduced his Three-Stratum Theory (TST), which was the result of a massive systematic exploratory factor analysis of essentially all extant correlation studies of mental test data and cognitive ability datasets.<sup>257</sup> Briefly, the TST is a three-tiered model of human cognitive abilities organized according to breadth. The broadest, top-most tier (stratum III) is a general intelligence factor (equivalent to Spearman’s *g*). Stratum II includes eight broad abilities that represent “basic constitutional and long-standing characteristics of individuals that can govern or influence a great variety of behaviors in a given domain.”<sup>258</sup> These include fluid intelligence (*Gf*), crystallized intelligence (*Gc*), general memory and learning (*Gy*), broad visual perception (*Gv*), broad auditory perception (*Ga*), broad retrieval ability (*Glr*), broad cognitive speediness (*Gs*), and reaction time/decision speed (*Gt*). Finally, stratum I includes 69 narrow abilities, each of which is subsumed by associated stratum II abilities, which in turn are subsumed by the single stratum III *g* factor. The reader will notice the small overlap with elements of the Cattell-Horn *Gf-Gc* theory; yet TST’s taxonomy is vastly more complex and, according to those in the field, unparalleled in the veracity of its conclusions. “In a sense, Carroll provided the field of intelligence a much needed *Rosetta Stone* that can serve as a key in

---

<sup>255</sup> John Horn, “Organization of abilities and the development of intelligence,” *Psych. Review* 75 (1968).

<sup>256</sup> The latter (*Gc*) subsumes Cattell’s earlier definition.

<sup>257</sup> Carroll started with a set of about 1,500 papers, from which he culled over 450 datasets that he used as the basis of his eventual theory (John Carroll, *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*, New York: Cambridge University Press, 1993). Carroll’s effort was so prodigious that more than few researchers consider the fruits of his labor to be on par with what Whitehead and Russell did for mathematics (with their *Principia Mathematica*) or what Mendeleev did for chemistry (with his construction of the periodic table of elements). See Kevin McGrew and J. Evans, “Internal and External Factorial Extensions to the Cattell-Horn-Carroll (CHC) Theory of Cognitive Abilities: A Review of Factor Analytic Research since Carroll’s Seminal 1993 Treatise,” *Carroll Human Cognitive Abilities (HCA) Project Research Report # 2*, July 2004.

<sup>258</sup> Carroll, *Human Cognitive Abilities*, p. 634.

deciphering and organizing the mass of human cognitive abilities structural literature that has accumulated since the days of Spearman.”<sup>259</sup>

And so, at last, we finally introduce the CHC framework, which effectively integrates the Cattell-Horn *Gf-Gc* theory and Carroll’s TST. Since it has close to 100 narrow abilities and 18 broad abilities (though both numbers periodically change as minor revisions are made constantly), the CHC taxonomy is complex enough that there is no single “correct” way of depicting it. Figure 16. Summary of the CHC cognitive abilities framework organizes it around four conceptual groupings (*motor abilities, perceptual processing, controlled attention, and acquired knowledge*). **Appendix H** contains a complete taxonomy.

Figure 16. Summary of the CHC cognitive abilities framework

Knowledge	<b>Comprehension-Knowledge</b> Language comprehension and general knowledge	<b>Gc</b>	LD: Language development, VL: Lexical knowledge, KD: General (verbal) information, LS: Listening ability, CM: Communication ability, MY: Grammatical sensitivity, K2: Information about culture, KI: Foreign language proficiency, LA: Foreign language aptitude
	<b>Domain-Specific Knowledge</b> Declarative and procedural knowledge related to specialized interests	<b>Gkn</b>	K1: General science information, K2: Knowledge of culture, MK: Mechanical knowledge, KI: Foreign language proficiency, KS: Knowledge of signing, LP: Skill in lip reading
	<b>Reading and Writing</b> Declarative and procedural knowledge related to literacy	<b>Grw</b>	RC: Reading comprehension, RD: Reading decoding, WA: Writing ability, RS: Reading speed, SG: Spelling ability, WS: Writing speed, EU: English usage
	<b>Quantitative Knowledge</b> Declarative and procedural knowledge related to mathematics	<b>Gq</b>	KM: Mathematical knowledge, A2: Mathematical achievement
	<b>Long-Term Storage and Retrieval</b> Store/consolidate new information and fluently retrieve stored information	<b>Gl</b>	MA: Associative memory, MM: Meaningful memory, MF: Free recall memory
	<b>Retrieval Fluency</b> The rate at which access information stored in long-term memory accessed	<b>Gr</b>	Fl: Ideational fluency, FE: Expressional fluency, LA: Speed of lexical access, NA: Naming facility, FW: Word fluency, FA: Associational fluency, SP: Alternative solution fluency, FO: Creativity, FF: Figural fluency, FW: Figural flexibility
Attention	<b>Fluid Reasoning</b> Use deliberate and controlled mental operations to solve novel problems	<b>Gf</b>	I: Induction, RQ: Quantitative reasoning, RG: Deductive reasoning, RE: Reasoning speed, RP: Piagetian reasoning
	<b>Short-Term Working Memory</b> Apprehend / be aware of information useful for multi-step problem solving	<b>Gwm</b>	WM: Working memory capacity, MA: Associative memory, MM: Meaningful memory, WA: Auditory short-term storage, WV: Visual-spatial short-term storage, AC: Attention control
	<b>Cognitive Processing Speed</b> Fluently perform relatively easy elementary cognitive tasks	<b>Gs</b>	P: Perceptual speed, Ps: Perceptual speed-search, Pc: Perceptual speed-compare, R: Number facility, RS: Reading fluency, WS: Writing fluency
	<b>Visual-Spatial Abilities</b> Perceive, discriminate, and manipulate images	<b>Gv</b>	Vz: Visualization, SR: Spatial relations, IM: Imagery, CS: Closure speed, CF: Flexibility of closure, MV: Visual memory, SS: Spatial scanning, PI: Serial perceptual integration, LE: Length estimation, IL: Perceptual illusions, PN: Perceptual alternations, P: Perceptual speed
	<b>Auditory Processing</b> Perceive, discriminate, and manipulate sounds	<b>Ga</b>	PC: Phonetic coding, UR: Judging Rhythm, US: Speech sound discrimination, LR: Resistance to acoustic distortion, UM: Memory for sound patterns, UI: Musical discrimination, UP: Absolute pitch, UL: Sound localization
	<b>Tactile (Haptic) Abilities</b> Perceive, discriminate, and manipulate touch stimuli	<b>Gh</b>	TS: Tactile Sensitivity
Perception	<b>Olfactory Abilities</b> Perceive, discriminate, and manipulate smells	<b>Go</b>	OM: Olfactory memory, OS: Olfactory sensitivity
	<b>Kinesthetic Abilities</b> Perceive, discriminate, and manipulate sensations of body movement	<b>Gk</b>	KS: Kinesthetic Sensitivity
	<b>Emotional intelligence</b> Perceive & understand emotional behavior; solve problems using emotions	<b>Gei</b>	Eg: Emotional perception, Ek: Emotion knowledge, Em: Emotion management, Eu: Emotion utilization
	<b>Reaction and Decision Speed</b> Speed at which very simple perceptual discriminations can be performed	<b>Gt</b>	R1: Simple reaction time, R2: Choice reaction time, IT: Inspection time, RM: Semantic processing speed, RZ: Mental comparison speed
Motor	<b>Psychomotor Abilities</b> Skilled performance of motor functions	<b>Gp</b>	P1: Manual dexterity, P2: Finger dexterity, P3: Static strength, PE: Cross body equilibrium, PC: Multilimb coordination, P7: Arm-hand steadiness, PB: Control precision, AI: Aiming
	<b>Psychomotor Speed</b> Speed of motor functions	<b>Gps</b>	PF: Speed of articulation, MT: Movement time, RL: Speed of limb movement, WS: Writing fluency

Source: CNA.

<sup>259</sup> McGrew and Evans, “Internal and External Factorial Extensions to the Cattell-Horn-Carroll (CHC) Theory of Cognitive Abilities,” p. 4.

In Figure 16, the center column contains CHC's stratum II abilities, which are color-coded according to *intelligence as knowledge* (in **black**), *intelligence as process* (in **blue**), intelligence as fluency or speed (in **green**), and (tentatively defined) *broad abilities* (in **pink**).<sup>260</sup>

## Toward a common language

The reader may be wondering why we have taken such care in presenting CHC's strata. Recall that this final section's narrative is working toward introducing a "template of a framework" to help *operationalize* military applications. As we will soon see, the utility of this template depends heavily on having a mutually translatable language that bridges AI/ML developers and practitioners (and the techno-conceptual milieu in which they work) on one side, and military stakeholders and decision-makers (and the tactical-strategic milieu in which they work) on the other. What we are proposing is that **both sides can benefit by using CHC as that bridge**, the operative assumption being that when either side engages the other (or when otherwise enmeshed in internal machinations regarding a project of mutual interest), each side is strongly motivated to structure their thoughts and deliberations as clearly and as unambiguously as possible (from the point of view of how a given side *expects the other to interpret their message*).

As we have stressed throughout this paper, AI-research-level reports and summaries are, at best, prone to oversimplification (and/or obfuscation) and, at worst, simply incomprehensible when presented in fully rigorous form. Unless stakeholders are sufficiently well technically trained, they cannot generally be expected to have an intuitive grasp of how, for example, "long short-term memory neural network architectures" *may* help in developing new heuristic search and targeting algorithms for an unmanned aerial vehicle (UAV); which *specific* ML methods are applicable to facilitating and improving the mission performance of offensive-cyber operations; or, in the context of adjudicating S&T portfolio options, which AI techniques are germane for, say, enhancing certain military tasks and operations. But if "messages" that contain these otherwise-difficult-to-understand *artificial* intelligence methods are first "translated" (via a common CHC-derived language) into more intuitively familiar terms and concepts that describe *human* intelligent abilities before they are sent, the military stakeholders that receive the "translations" stand a far better chance of understanding their intended meaning.

Of course, we are not suggesting a literal translation. But we are strongly advocating that what is urgently needed is a framework that facilitates a mutually understandable dialogue. On the military side, the lone, still-missing ingredient is a way to translate military operational

---

<sup>260</sup> The CHC abilities that appear in Figure 16 and **Appendix H** are taken from the latest taxonomy (v2.5). See Kevin McGrew, "Cattell-Horn-Carroll (CHC) theory of cognitive abilities (v2.5) 'official' broad and narrow definitions," IQ's Corner, 10 July 2017, <http://www.iqscorner.com/2017/07/cattell-horn-carroll-chc-theory-of.html>.

concepts (and tasks and missions) into terms that non-militarily-trained stakeholders (i.e., typical AI/ML researchers) can “understand.” Luckily, such a device already exists and is well known to the military community: the OODA loop.

## Observe-orient-decide-act (OODA)

Figure 17 shows a schematic view of a basic OODA loop overlaid with elements relevant to both AI/ML and military tasks;<sup>261</sup> the *Orient*→*Decide* parts are highlighted (in blue) to emphasize that many of their embedded functions overlap with “cutting-edge” developments and advances in AI/ML. The OODA loop is a simple model of decision-making introduced by USAF Col. John Boyd in the late 1960s.<sup>262</sup> Intended originally as a conceptual backdrop to facilitate discussion and analyses of air combat (and military strategy in general),<sup>263</sup> it has since been applied to widely diverse fields that involve decision-making in adversarial environments (e.g., business, law enforcement, and sports).<sup>264</sup> The OODA loop has also been used to model planning and human supervisory control of physical systems.<sup>265</sup>

The OODA loop’s seeming “simplicity” (at least as depicted in Figure 17) is deceptive. For one thing, Boyd’s original design was considerably more complex and included additional subprocesses and information flows.<sup>266</sup> For another, it masks an overarching rubric grounded in dynamic control theory and the rudiments of cybernetics (a precursor to complex systems and AI itself).<sup>267</sup> In other words, the OODA loop essentially represents an autopoietic process, where the internal operations of the human—or, in our context, *human-AI hybrid*—interact with, and adapt to, external changes in the environment.

---

<sup>261</sup> Figure 17 is based on the 2017/AI paper, Figure 29 (p. 157), which also contains a more detailed discussion of the OODA loop than appears here.

<sup>262</sup> R. Coram, *Boyd: The Fighter Pilot Who Changed the Art of War*, New York: Back Bay Books, 2004.

<sup>263</sup> 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-15s were technically superior to the American F-86 Sabres, Boyd argued that it was because of the F-86’s superior cockpit visibility that US pilots were able to decide and act faster than their opponents; that is, US pilots were generally able to get into a good firing position before their Korean counterparts could react. See F. Osinga, *Science, Strategy, and War*, New York: Routledge, 2006.

<sup>264</sup> G. Hammond, *The Mind of War: John Boyd and American Security*, Washington, DC: Smithsonian Books, 2004.

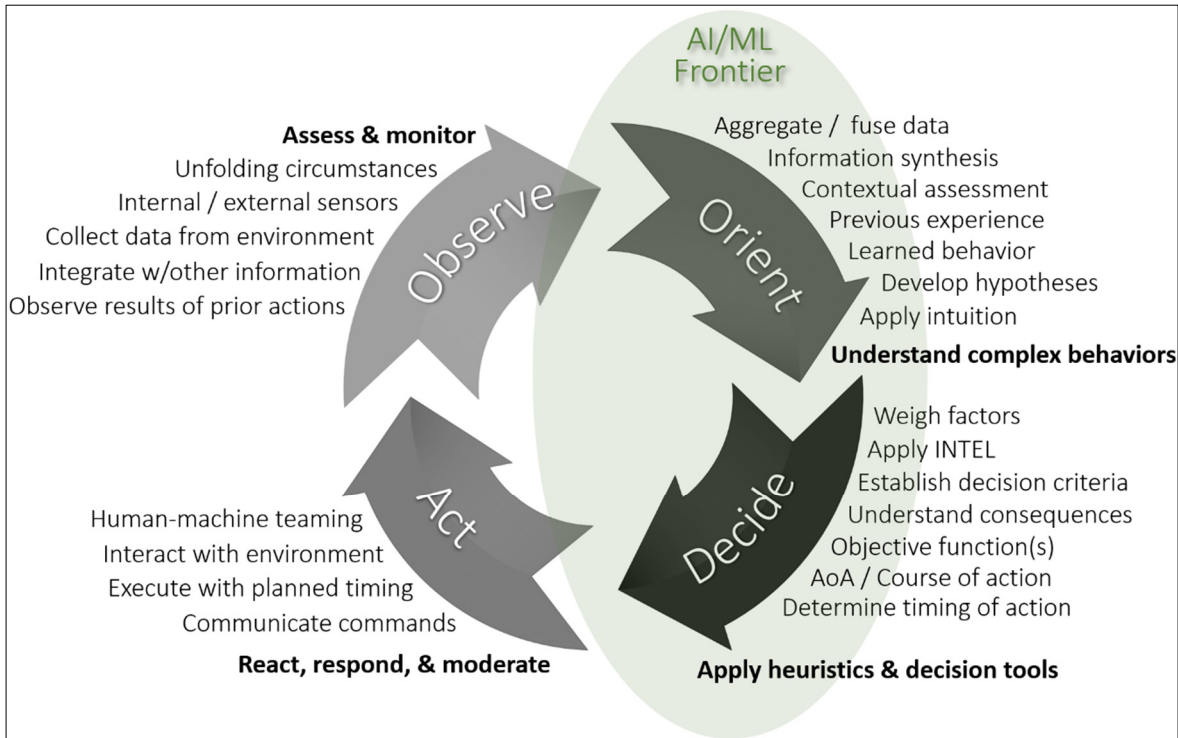
<sup>265</sup> T. Grant, “Unifying Planning and Control using an OODA-based Architecture,” *Proceedings of SAICSIT*, 2005.

<sup>266</sup> John Boyd, “The Essence of Winning and Losing,” US Army Command and General Staff College, June 1995, <https://web.archive.org/web/20110324054054/http://www.danford.net/boyd/essence.htm>.

<sup>267</sup> Norbert Wiener, *Cybernetics*, 2nd ed., Cambridge, MA: MIT Press, 1965.



Figure 17. OODA loop and elements pertaining to AI and ML



Source: CNA.

The four stages of the OODA loop’s cyclical, dynamic process are (1) *Observe*, which 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; (2) *Orient*, which consists of aggregating, correlating, and analyzing collected information and compiling a real-time situational awareness picture; (3) *Decide*, which involves weighing various factors and options against the assigned (and/or locally determined) objectives to determine a course of action; and (4) *Act*, which consists of following through on the decision (e.g., striking a target, applying navigational course correction, or engaging radar jamming). An implicit takeaway is that the OODA loop is not static but, rather, consists of multiple interlinking processes with entwined inputs and decision steps. It is a continuous loop and involves myriad simultaneous decisions that must be adjudicated in parallel. Information, too, is fluid and 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.<sup>268</sup>

<sup>268</sup> R. Deakin, *Battlespace Technologies*, Boston: Artech House, 2010.

Another implicit takeaway from Figure 17 (see the portion within the green-shaded oval) is that AI-infused autonomous systems perform an analogous set of functions to what humans require to accomplish given tasks. While certain parts of the *Observe* and *Act* phases may be straightforward “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), the functions that make up the *Orient* and *Decide* phases of the OODA loop contain many key AI/ML-driven and other computation-based capabilities. *Orient* involves functions that determine how well a system is able to “understand” its environment: aggregating and fusing asynchronous data from multiple data sources; hypothesizing about and deducing features to describe 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, and general “perception” algorithms that are able to fuse and draw context-relevant inferences from multiple forms of sensory input are still in nascent form as of this writing (September 2020).

The *Decide* stage 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 anticipating adversarial countermeasures in the next *Decision* cycle. Many of these elements are, for those just recounted for the *Orient* stage, at the cutting edge of current 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.

As a precursor to the CHC-based “template of a framework” introduced in the next section, consider the OODA loop-based taxonomy of *autonomy* (viewed, loosely, as an assembly of AI-infused subsystems) shown in Figure 18.

Before moving on to the final section, we wish to highlight this taxonomy’s elegant enfolding of *one* large multidimensional space—consisting of operational tasks and missions (along with a sampling of required skills)—into *another*; that is, the myriad actions and processes subsumed in the OODA loop’s four main stages.

Figure 18. OODA loop–based autonomy taxonomy

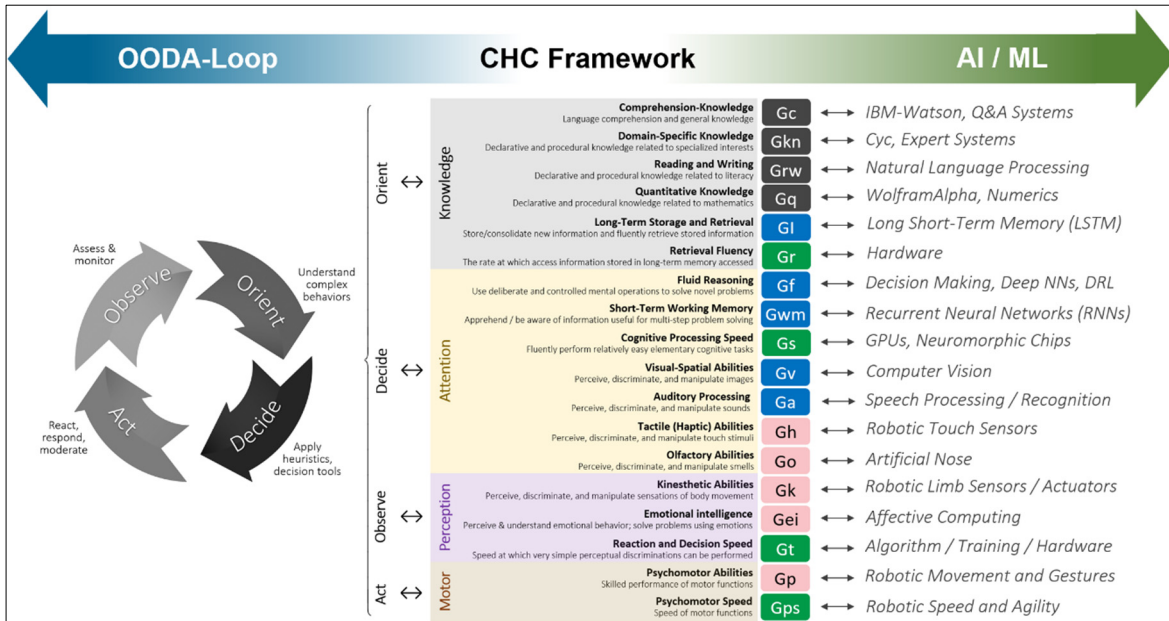
Level	Description	Observe	Orient	Decide	Act
		Perception / Situational Awareness	Analysis / Coordination	Decision Making	Capability
0	Remotely piloted vehicle	Flight control (attitude, rates) sensing; on-board camera	Telemetered data; remote pilot command	N/A; off-board pilot	Control by remote pilot
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

Source: E. Sholes, “Evolution of a UAV Autonomy Classification Taxonomy,” *IEEE Aerospace Conference*, 2007.

## One possible bridge to help DOD pave a path from “understanding” to operationalizing AI

This final section introduces a “template of a framework” designed to help bridge the gap between “understanding AI” and *operationalizing* its military applications. We call it a “template” and not a full framework because it is precisely that—namely, a concept for what a fully developed framework might ultimately look like. As such, it is intended only to whet the appetite of interested stakeholders (AI/ML researchers and military stakeholders alike) and to jump-start a brainstorming process that is still needed to fill in missing parts and details. But the concept is at least a well-informed one, since it leverages knowledge bases that span all three pertinent domains (see Figure 19): *artificial intelligence*, *military operations*, and—via *CHC*—*human intelligence and cognition*.

Figure 19. Toward enfoldng AI/ML, the OODA loop, and the CHC cognitive framework



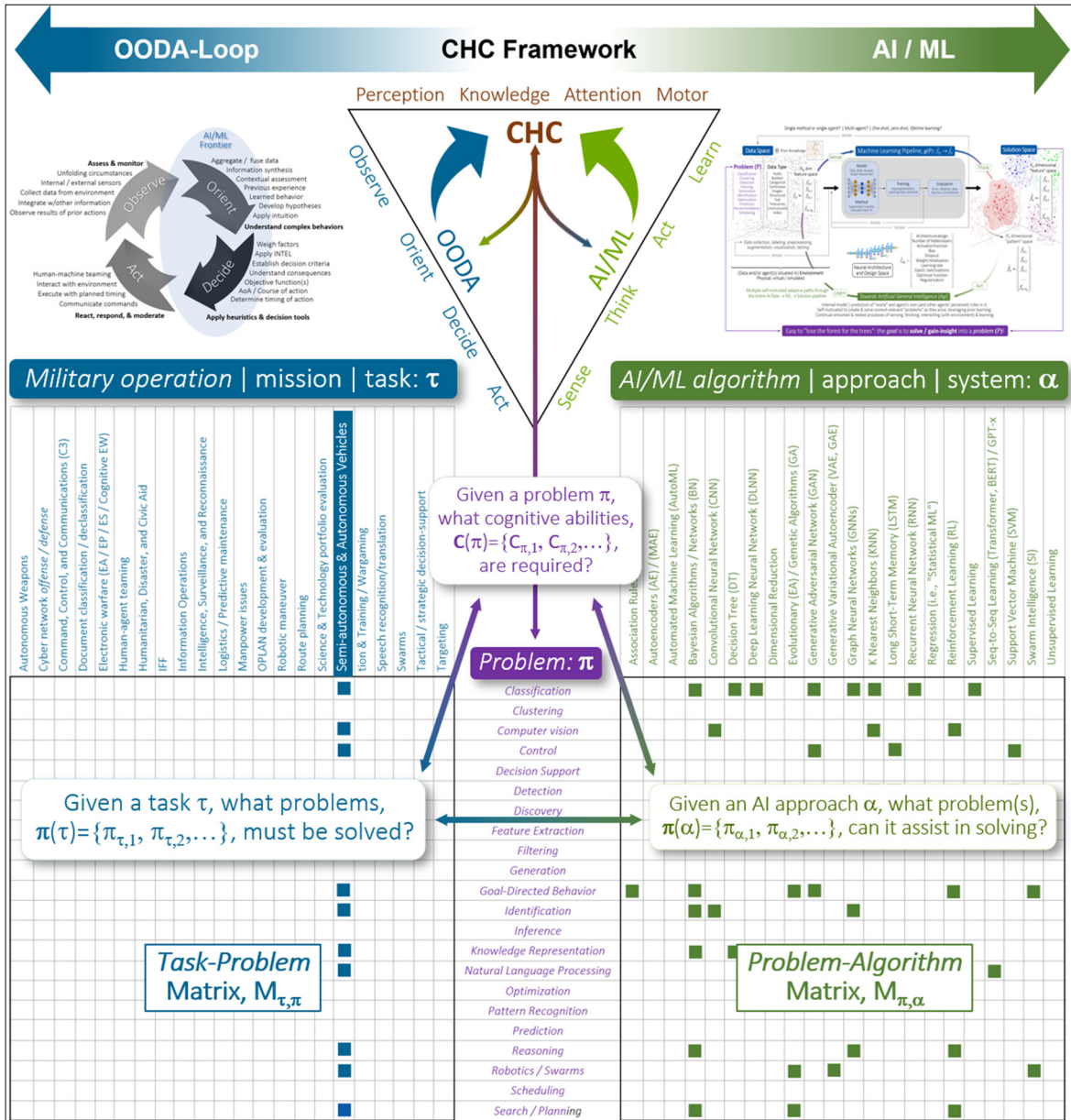
Source: CNA.

The left-hand side of Figure 19 shows the OODA loop as the key that relates the sets of skills and abilities required of *any intelligent system* (human, AI, and/or hybrid, and as defined by the CHC taxonomy) in order to perform specific military functions, tasks, and missions. Note the natural association between the OODA loop’s *Orient–Decide–Observe–Act* foundation and CHC’s *Knowledge–Attention–Perception–Motor* decomposition. The right-hand side shows a “quick-and-dirty” sample mapping between CHC’s mid-tier stratum II abilities and relevant algorithms and methods that live in the “AI/ML part” of the spectrum. The last remaining step is to leverage CHC as the foundation that explicitly links military tasks and AI methodology.

### Final assembly of the military, AI, and CHC taxonomy jigsaw pieces

Figure 20 shows the schematic assembly of all of the relevant pieces of our jigsaw puzzle; that is, the “template of a framework” for a bridge—or *mutually translatable language*—between military stakeholders and decision-makers, on one side, and AI/ML developers and practitioners, on the other.

Figure 20. A concept for a CHC-mediated bridge between military operations and AI methodology



Source: CNA.

The final assembly is constructed around a set of fundamental questions that may be phrased, if only formally (for now) in terms of three factors, symbolically denoted by: (1)  $\tau$ , which represents military operations, missions, and/tasks (highlighted in blue); (2)  $\alpha$ , which symbolizes AI and ML methods and algorithms (highlighted in green); and (3)  $\pi$ , which connotes a set of general problems that straddle both military and AI/ML domains (highlighted in purple). The CHC cognitive abilities framework—highlighted in brown at the top of the

upside-down triangle—serves as the central conceptual and dynamic pillar of the whole assembly. Color-coded incoming and outgoing arrows represent specific functions administered by the CHC: (**blue**) OODA-loop-derived military tasks are “translated” into an AI/ML-grounded “language” before flowing back to the military side of the spectrum; (**green**) AI/ML-derived methods and algorithms are “translated” into an OODA loop–derived “language” before flowing back to the AI/ML side of the spectrum; and the two-sided downward arrow (**brown** on top, and **purple** on the bottom) connects the CHC cognitive abilities taxonomy with problems common to both military and AI/ML sides of the dialectic spectrum.

The list of (alphabetically ordered) problems that appears in the middle column in the bottom half of Figure 20 represent typical applications of specific AI/ML methods and algorithms (e.g., *classification, clustering, computer vision, control, decision support*, etc.). A given problem, say, “classification,” may be amenable to a “solution” using one or more AI/ML classes of algorithms (e.g., Bayesian, decision trees, and deep-learning networks). The *Problem-Algorithm Matrix*,  $M_{\pi,\alpha}$  to the *right* of the list of problems makes this association explicit: if an algorithm in the  $\alpha^{\text{th}}$  column is appropriate for tackling a problem in the  $\pi^{\text{th}}$  row, a solid **green** block, ■, appears in the  $(\pi,\alpha)^{\text{th}}$  site of the matrix. Similarly, the *Task-Problem Matrix*,  $M_{\tau,\pi}$  to the *left* of the list of problems encodes the types of problems that a given military task entails the solution of: if a task in the  $\tau^{\text{th}}$  column requires “solving” a problem in the  $\pi^{\text{th}}$  row, a solid **blue** block, ■, appears in the  $(\tau,\pi)^{\text{th}}$  site of the matrix. Implicit in the assembly (and not shown in Figure 20) is a third matrix—specifically, the *CHC-Problem Matrix*,  $M_{C,\pi}$ —which encodes the types of cognitive abilities relevant for solving a given “problem.”<sup>269</sup>

Collectively, these matrices facilitate a mutual exploration among *all* stakeholders (spanning the entire military-centric  $\leftrightarrow$  AI/ML-centric spectrum) of three complementary sets of questions:

- Given a task  $\tau$ , what kinds of problems,  $\pi(\tau)=\{\pi_{\tau,1}, \pi_{\tau,2},\dots\}$ , must be solved?
- Given an AI/ML approach,  $\alpha$ , what kinds of problems,  $\pi(\alpha)=\{\pi_{\alpha,1}, \pi_{\alpha,2},\dots\}$ , can it assist in solving?
- Given a problem,  $\pi$ , which general cognitive abilities,  $C(\pi)=\{C_{\pi,1}, C_{\pi,2},\dots\}$ , are relevant in finding a solution(s)?

Of course, it requires a tremendous effort to systematically and credibly fill in all entries for all three matrices, which is a primary reason for calling this only a *template* of a framework. It is not trivial to even define the rows and columns of these matrices, let alone the values of

---

<sup>269</sup> Adding another matrix would have made it difficult to render visible all elements of the assembly in a single page-size graphic. Its implicit form is analogous to the two matrices that do appear. Namely, if a CHC cognitive ability,  $C$ , is relevant for solving a problem,  $\pi$ , then a solid block, ■, appears in the  $(C,\pi)^{\text{th}}$  site of the matrix.

individual entries (i.e., whether a given row element  $x$  is associated with a specific column entry  $y$ ). For example, the 22 military tasks that appear in the columns of the task-problem matrix (highlighted in **blue** on the left-hand side of Figure 20) make up but a tiny fraction of the “military operational space.” **Appendix I** contains a larger, but still incomplete taxonomy and gives an idea of the difficulty that just this one part entails before the caveat “template of...” can be removed from the framework. Similarly, the 22 AI methods populating the columns of the problem-algorithm matrix (highlighted in **green** on the right-hand side of Figure 20) are but a small sampling of a vastly larger space of methods and techniques (with new ones being spawned seemingly every other week). **Appendix E** provides a larger taxonomy.

Nonetheless, even without a fully developed framework to work with, the utility of having such a framework is clear. Consider the notional example that appears in Figure 20—namely, the “task” that consists of developing a semi-autonomous vehicle (SAV), which appears as the lone column in the task-problem matrix whose associated row entries are identified. An SAV’s inherent complexity (as well as that of the larger ecosystem it is a part of once it deploys) entails “solving” many broad problems that are in principle amenable to AI/ML methodology; for example (among many others than can be listed):<sup>270</sup>

- *Perception*—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*—generating a map of a 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 impede the SAV’s progress.
- *Planning*—generating a sequence of actions to take from a specified starting position to its final destination (or activity) while avoiding obstacles and other “unanticipated” impediments.
- *Behavior*—translating the combined outputs of the navigation, planning, and perception functions into actuator commands that allow the SAV to execute specific actions (e.g., move and/or fire weapons).
- *Targeting*—adjudicating which targets to engage (which also requires a contextual consistency with an SAV’s planning and behavior modules).

Each of these broad problems, in turn, entails solving more focused problems (i.e., the  $\pi$ -type problems in the lower center column of Figure 20); specifically, those for which the (SAV, $\pi$ ) entries contain a **blue** block: *classification, computer vision, control, detection, feature*

---

<sup>270</sup> Chapter 3 in A. Finn and S. Scheding, *Developments and Challenges for Autonomous Unmanned Vehicles*, New York: Springer-Verlag, 2010.

extraction, and so on. The (green block) entries of the *problem-algorithm* matrix on the right allow us to decompose—and translate an SAV’s developmental requirements—still further in terms of basic AI/ML methodology. For example, *association rules*, *Bayesian algorithms*, and *evolutionary programming methods* are all applicable to “solving” the goal-directed behavior problem.

If the only goal is to assess the relative merits of alternative AI investment options (as part of an S&T portfolio analysis) for *existing* tasks or systems—or capabilities for which the military already has well-understood “categories,” such as the SAV just discussed—the *problem-algorithm* and *task-problem* matrices suffice to inform this decision process; in other words, there is no need to involve the CHC framework. Although, even in this case, options must still be weighed according to, for example, the level of maturity of an AI method, its adoptability, cost of additional development that may be required, availability of suitable datasets, reliability, vulnerability to exploitation and/or attack, etc.

But if the goal is to discover and explore innovative *new* AI options for heretofore unrecognized, unknown, or “untapped” tasks, mission capabilities, and/or systems, the overarching CHC framework can be used as a direct go-between for military stakeholders and the AI research community; that is, as the foundation on which the potential *operationalizability* of AI methods (including AI-human hybrid systems) is assessed through *psychometric profiling*.

“Psychometric profiling” refers to the practice of weighing all pertinent dimensions of an applicable set of behavioral features (i.e., deliberately *not* reducing measures of, say, “intelligence” to a single number).<sup>271</sup> But while the general technique is not new, it has only recently appeared in discussions of machine intelligence.<sup>272</sup> We propose that its utility may run far deeper: specifically, CHC’s already comprehensive taxonomy may be leveraged to define multidimensional profiles of “general cognitive abilities” that once accomplished can, in turn, be used to inform both sides of the OODA loop ↔ AI/ML spectrum as they engage in a mutual dialectic aimed at discovering innovative systems and approaches (the figurative takeaway from the two-sided arrow that connects the “CHC Framework” at the top of Figure 21 to the list of problems,  $\pi$ , highlighted in purple below).

Figure 21 shows a few notional CHC-derived psychometric profiles for a typical “human intelligence” and three hypothetical “AI intelligences.” Where humans are expected to demonstrate strong correlations among a broad set of basic skills and abilities, “AI intelligences” are unburdened by such a priori constraints. Some may possess strong abilities

---

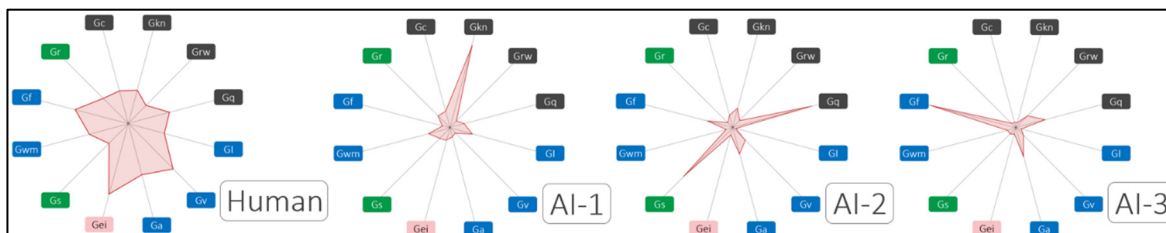
<sup>271</sup> E. Messina, A. Meystel, and L. Reeker, “Measuring Performance and Intelligence of Intelligent Systems,” PERMIS 2001 White Paper, NIST, [https://www.nist.gov/system/files/documents/el/isd/ks/WhitePaper\\_2001.pdf](https://www.nist.gov/system/files/documents/el/isd/ks/WhitePaper_2001.pdf).

<sup>272</sup> Graziell Orru et al., “Machine Learning in Psychometrics and Psychological Research,” *Frontiers in Psychology* 10 (January 2020), <https://doi.org/10.3389/fpsyg.2019.02970>.



for  $x$ ,  $y$ , and  $z$ , and be poor at all the rest; others may be organized in clusters of complementary abilities, wherein one group is good half of a given stratum II skill set, and the other good at the other half; and still others may be highly specialized (i.e., uniquely strong performers) in only one or two skills from different strata, and for which there is no direct human analog.

Figure 21. Notional examples of CHC-based psychometric profiles of "general intelligence"



Source: CNA.

Coming (almost) full-circle, recall our earlier depiction of the history of AI, ML, and deep learning neural architectures as a creative, self-organized evolutionary traversal of an abstract multidimensional space (see **“The ‘adjacent possible” in an AI/neural-LEGO World”**): we conclude this paper with one last “template” of a proposal—that *AI itself be used as a creative exploratory tool* to help military stakeholders (analysts, S&T portfolio analysts, and CONOPS developers) navigate the vast space of the *adjacent possible*. The landmark achievements of AI systems such as AlphaZero, AlphaStar, and GPT-3, and that of other recent algorithms (see **“AI ‘hits’ during 2017–2020”**) all demonstrate a superhuman ability to find “solutions” in what are, for humans, unimaginably large search spaces. To emphasize: they are not just superhuman performers on the narrow tasks they were each developed to “solve” (chess, shogi, and Go, in the case of AlphaZero; StarCraft II, in the case of AlphaStar; and text generation, in the case of GPT-3); they also all share a superhuman ability to creatively traverse immensely complex multidimensional abstract spaces.<sup>273</sup> This innate capacity for creative exploration can, in principle, be leveraged to help human stakeholders discover innovative approaches to “old” problems and/or discover heretofore unknown solutions to problems not yet recognized.

<sup>273</sup> Risto Miikkulainen, “Creative AI Through Evolutionary Computation,” 22 Feb. 2020, arXiv:1901.03775v2; Marian Mazzone and A. Elgammal, “Art, Creativity, and the Potential of Artificial Intelligence,” *Arts* 8 (2019); Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity*, Cambridge, MA: MIT Press, 2019.

[This page intentionally left blank]

# Summary and Conclusions

---

This paper is intended to be a direct sequel to CNA's *AI, Robots, and Swarms* report, published in 2017, in order to bring to bring the summary and conclusions of that earlier report up to date by incorporating more recent AI-related milestones and developments.

On the one hand, not much has changed, in the sense that AI's exponential growth in the academic and commercial research communities continues unabated; indeed, there are signs that it is accelerating. The most stunning development reported in the 2017 paper—namely, AlphaGo's defeat of 18-time world Go champion Lee Sedol in 2016—has, remarkably, seemingly faded into history, as the historic feat was superseded not once but twice in less than two years: first, in December 2017, AlphaZero, starting from random play and using no domain knowledge except for game rules, required only 24 hours to achieve a superhuman level of play in chess, shogi, and Go (and defeated a world-champion program in each), and then in November 2019, MuZero matched AlphaZero's superhuman performance without any knowledge of game rules! DOD's unclassified investments in AI have increased from about \$600 million in FY 2016 to \$927 million in FY 2020, with more than 600 active AI-related projects (as of August 2019). DOD has requested \$800 million in FY 2021 to continue "the AI pathfinders, JAIC, and advanced image recognition, and an additional \$1.7 billion for autonomy." And DARPA continues to push the technological envelope by announcing the establishment of, or further developments in, at least 40 different basic research programs (between the beginning of 2017 and mid-2020). Just as the final pages of this report were being written, DARPA announced on 20 August 2020 that an AI "pilot" developed under its Air Combat Evolution (ACE) program, launched in May 2019, defeated an Air Force F-16 piloted by a human in a simulated aerial dogfight contest.

On the other hand, signs of trouble are unmistakably brewing. There remains, unfathomably, little or no consensus about what *AI means*, either among the AI research community at large or within DOD and other US government agencies. While this by itself is not an immediate cause for alarm, a lack of a universally agreed-upon definition is nonetheless troubling because stakeholders have no framework on which to base S&T investment strategies or policy deliberations. The lack of a definition is also potentially but the first in a Pandora's Box of related concerns, starting with the continued divide between military stakeholders and decision-makers, on one side, and AI developers and practitioners, on the other. As AI methodology becomes increasingly opaque and difficult to understand, even for developers (e.g., while the research community struggles to reduce the "explainability gap" to better understand the behavior of deep-learning systems whose architectures were designed by humans, AutoML systems are now designing systems whose *architectures* are also difficult to

understand), it is becoming increasingly difficult for military stakeholders in authoritative decision-making positions to keep pace, particularly stakeholders not sufficiently conversant to keep abreast of the last techniques and algorithms.

The list of fundamental gaps, limitations, and challenges reported in the 2017 paper not only remains mostly intact (few, if any entries, are officially “off the list”), but also appears to be growing. For example, while it was already known in 2017 that deep learning neural networks are prone to “adversarial attacks” (i.e., they all 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), the list of algorithms and the basic vulnerabilities they each exploit have grown considerably since: “Single-pixel,” “adversarial patch,” “elephant in the room,” “poison frog,” and “energy latency” adversarial attacks are but several of the many methods that have been introduced just in the last two years alone (a recent survey lists more than 30).

There is evidence to suggest that while we may not quite be headed for another “AI winter” (AI has previously weathered at least two major winters during its six decade-plus history), we may have at least entered, or are soon to enter, an era of diminishing returns. One recent study concludes that continued progress “will require dramatically more computationally efficient methods, which will either have to come from changes to deep learning or from moving to other machine learning methods.” For example, the conventional wisdom only a few years ago was that self-driving cars are “just around the corner.” The reality, as one recent MIT study found (see section, “**New AI ‘Challenges’**”), is that the “the full driving task is too complex an activity to be fully formalized as a sensing-acting robotics system that can be explicitly solved through model-based and learning-based approaches.”<sup>274</sup>

There are also important lessons to be drawn from how the AI research community has responded to the ongoing (as of this writing, September 2020) COVID-19 pandemic.<sup>275</sup> It is sobering to acknowledge the unequivocal agreement of the dozen or so major reviews of this response that were consulted for this study that collectively examined more than a thousand published research papers and methodological techniques: virtually all COVID-19 and AI-related studies are challenged by a lack of available large-scale training data, datasets prone to high risk of bias, massively “noisy” data, and the general propagation of misinformation and unverified rumors on social media sites; a knowledge gap between AI experts and other computer scientists and medical ethics-related gaps (e.g., data privacy and human rights protection), rapid-fire-deployment without proper peer-review; and a high propensity to report extremely optimistic results that are not warranted by actual results.

---

<sup>274</sup> L. Fridman et al., “MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction with Automation,” *IEEE Access* 7, 1 July 2019.

<sup>275</sup> Juan Mateos-Garcia, J. Klinger, and K. Stathoulopoulos, *Artificial Intelligence and the Fight Against COVID-19*, NESTA, 15 June 2020, <https://www.nesta.org.uk/report/artificial-intelligence-and-fight-against-covid-19/full/>.

The AI community’s response to COVID-19 demonstrates, in textbook fashion, that substantively applying AI to a problem that lies *outside the scope and expertise of typical ML researchers is hard, very hard*—particularly when relevant datasets are unavailable (or are in short supply, not sufficiently representative of the requisite “training” space, and/or biased)—and succeeds best when researchers and problem-domain experts (in COVID-19’s case, medical practitioners, and in DOD’s case, military leaders, policy-makers, and warfighters) are all aligned with common goals, priorities, methodology, and—above all else—a *common language*.

The penultimate section of this paper introduces a “template of a framework” for developing this common language. We call it a “template” because its current embryonic form only hints at what a fully developed framework might eventually look like. The conceptual assembly of this template depends critically on our going (almost) full circle by asking not what “artificial intelligence” is (thus avoiding the quagmire discussed in the first part of the paper) but, rather, what “intelligence” is, generally, regardless of its origin (human cognition, AI, or some as-yet-unrealized human-AI hybrid). The resulting framework combines precepts and taxonomies for AI, military tasks and operations, and general human intelligence and cognition. When fully realized, it can be used to bridge the stubbornly persistent gap between “understanding AI” and fully *operationalizing* its military applications.

## Recommendations

We began this narrative by recounting how Section 238 of the FY 2019 National Defense Authorization Act, released in August 2018, called for<sup>276</sup> (1) the establishment of the JAIC to coordinate DOD projects of more than \$15 million; (2) DOD to publish a strategic roadmap for AI development and deployment (which was published in February 2019);<sup>277</sup> and (3) the Secretary of Defense to produce a definition of AI by 13 August 2019 (which remains undone as of this writing in September 2020). Yet the 2019 NDAA also includes a set of broad “recommendational elements” meant to guide a more detailed implementation of the 2018 AI Strategy plan. These elements include the following:<sup>278</sup>

*Element 3-A: A comprehensive and national-level review of—*

- advances in artificial intelligence, machine learning, and associated technologies relevant to the needs of the Department [of Defense] and the Armed Forces; and

---

<sup>276</sup> John S. McCain National Defense Authorization Act for FY19, Public Law 115-232, 13 Aug 2018. <https://www.congress.gov/115/plaws/publ232/PLAW-115publ232.pdf>

<sup>277</sup> Summary of the 2018 DOD AI Strategy: <https://fas.org/man/eprint/dod-ai.pdf>.

<sup>278</sup> Pub. L. 115-232, 132 Stat. 1697.

- the competitiveness of the Department in artificial intelligence, machine learning, and such technologies.

*Element 3-B:* Near-term actionable recommendations to the Secretary [of Defense] for the Department to secure and maintain technical advantage in artificial intelligence, including ways—

- (i) to more effectively organize the Department for artificial intelligence;
- (ii) to educate, recruit, and retain leading talent; and
- (iii) to most effectively leverage investments in basic and advanced research and commercial progress in these technologies.

*Element 3-C:* Recommendations on the establishment of Department-wide data standards and the provision of incentives for the sharing of open training data, including those relevant for research into systems that integrate artificial intelligence and machine learning with human teams.

*Element 3-D:* Recommendations for engagement by the Department with relevant agencies that will be involved with artificial intelligence in the future.

*Element 3-E:* Recommendations for legislative action relating to artificial intelligence, machine learning, and associated technologies, including recommendations to more effectively fund and organize the Department.

We conclude this paper by offering five recommendations for specific actions and future studies that address some of the concerns raised in the 2019 NDAA’s broader set of “recommendational elements”:

**Recommendation #1:** *Move away from myopically “simple,” static definitions of AI toward active engagement—and continual reengagement—of all stakeholders to adapt, adopt, and reconceptualize AI (as new technologies and methods are inevitably introduced) as a holistic Sense → Think → Learn → Act evolutionary cyclic process.* Among the DOD Inspector General’s (DODIG) recommendations (to JAIC, issued as part of its recent audit) is to “include a standard definition of AI and regularly, at least annually, consider updating the definition.”<sup>279</sup> Yet as this paper shows, definitions are unable to capture any meaningful undercurrent of what AI *really is*, a deficiency that is only exacerbated by the ongoing clash of stove-piped views stemming from various DOD (and other government) agencies and the general dissonance even within the AI research community. Working toward achieving a consistent appreciation and understanding of the elements in figures Figure 9 and Figure 20 (among all stakeholders) is a good start.

**Recommendation #2:** *Embrace the irreducible reality of the inherent challenges associated with developing and deploying AI systems and develop a set of formal practices and procedures*

---

<sup>279</sup> *Audit of Governance and Protection of Department of Defense Artificial Intelligence Data and Technology*, DOD Office of Inspector General, DODIG-2020-098, 1 July 2020.

for mitigating fundamental challenges. Apart from the persistent scholarship issues that afflict all basic AI research (reproducibility, replicability, dataset bias, etc.), the testing and evaluation (T&E) and VV&A of AI-infused military systems entails myriad layers of complexity, which become ever more egregious as systems give way to nonlinearly coupled systems-of-systems.<sup>280</sup> There is a critical need to develop and advance the underlying science for new AI-targeted testing and evaluation (T&E) and verification, validation, and accreditation (VV&A) practices, including new standards, new guidelines, and engineering practices. A landmark set of standards that has recently appeared in a nonmilitary domain—namely, *medical and health care technology*—offers strict, well-defined guidelines for conducting and reporting clinical trials that involve AI.<sup>281</sup> It is an alluringly suggestive benchmark for developing an analogous set of standards for T&E and VV&A of military AI systems.

**Recommendation #3:** *Anticipate, and thus mitigate, the otherwise bleak consequences of ignoring the possibility that deep learning may already have entered (or is soon to enter) an era of diminishing returns, in which ever-increasing computational resources and/or refinements in algorithmic techniques yield relatively marginal gains in performance.* This entails a looming transitional period during which DOD’s AI S&T portfolio investment strategies must shift from leveraging “low-hanging fruit” applications (e.g., using PyTorch or TensorFlow to help intelligence analysts parse satellite imagery) to focusing more on operationalizing (and/or adopting) existing state-of-the-art (SOTA) AI/ML methods (whose performance is already “good enough”). A “minimal-viability” approach was recently introduced to help address this issue (based on mapping AI performance to military utility),<sup>282</sup> but the deeper underlying problem remains unsolved. The ML community has come up with tools that help developers identify the “sweet spot” of, say, a neural network’s training regimen’s learning curve (i.e., that help decide whether a model is *under-fitting* or *over-fitting* a dataset, or is “well fit”),<sup>283</sup> but an analogous toolset for AI/S&T portfolio investment managers has yet to be developed.

**Recommendation #4:** *Develop a framework to better enable, foster, and nurture collaborative engagements—and a mutual dialectic—among AI researchers, technology developers, and military policy-makers, S&T portfolio managers, and warfighters.* This paper contains the “template” for one such approach (see “**Toward a common language**”), designed to help bridge the gap between “understanding AI” and operationalizing its military applications; but

---

<sup>280</sup> See discussions on pp. 94–95 and 199–204 in Ilachinski, *AI, Robots, and Swarms*. See also Brian Haugh, D. Sparrow, and D. Tate, *The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems*, Institute for Defense Analyses, IDA Paper P-9292, September 2018.

<sup>281</sup> Xiaoxuan Liu et al., “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension,” *Nature Medicine* 26 (9 Sept. 2020).

<sup>282</sup> Courtney Crosby, “Operationalizing Artificial Intelligence for Algorithmic Warfare,” *Military Review*, July-August 2020.

<sup>283</sup> “Under-fitting” refers to models unable to learn the training dataset; “over-fitting” occurs when a model learns the training dataset too well, including noise or random fluctuations in the training dataset, and is unable to generalize.

even a partially realized “minimally viable” version entails considerable effort. However, absent such a fully realized framework, all stakeholder communities face the specter of an increasingly growing dissonance—not just in regard to *defining* AI, but in reaching a mutual consensus of *how to best develop and deploy it*.

**Recommendation #5:** *Leverage AI’s innate superhuman capacity to search through vast, complex, multidimensional, abstract spaces to develop a radical, new class of ML tools to help human stakeholders discover innovative approaches to “old” problems and/or discover heretofore unknown solutions to problems not yet recognized.* For example, as DOD looks beyond the “low-hanging fruit” of AI applications (see **Recommendation #3**), it will be increasingly challenged to imagine entirely new *categories* of deploying AI. To date, military applications of AI have been confined mostly to those that enhance existing human capabilities (i.e., enabling faster, longer, more accurate levels of performance). But what of future “creative” possibilities heretofore untapped? Recalling Kauffman’s “unanticipatable” space of the “adjacent possible” (see **“The ‘adjacent possible’ in an AI/neural-LEGO World”**), multiple simultaneous AI abilities merge to form entirely new “systems of systems” for which there are, as yet, no categories in which they can be readily placed.

It is easy to imagine using (or imagining the need to develop) an “AI” designed to augment and/or replace some task currently performed by humans (e.g., Project Maven’s use of ML algorithms to assist human analysts with predictive maintenance). It is slightly more difficult, but it still requires no particularly special creative spark, to imagine applying AI to tasks that no human can perform alone, but for which a category already exists (e.g., AI-enabled sensor grids that “see,” “detect,” and “react to” far better and faster than any human). But it is all but impossible to imagine the vastly larger set of possible—and heretofore unknown—AI functions, tasks, and systems that live in the space of the adjacent possible. Just as tanks, GPS-guided missiles, and digital computers were all “inconceivable” before they were invented, AI will inevitably spawn new systems, capabilities, and mission sets that do not yet exist. And we have not even discussed the just-as-inevitable entwinement of AI with the Internet-of-Battle-Things (IoBT), quantum computation, and brain-computer interfaces (to name just a few newly emerging technologies). But AI’s ouroboric capacity for “creative search” can easily be used to jump-start the discovery of its own “as yet unimagined” applications.

Happily, the *technology* to develop and leverage such tools—if not the *tools* themselves—already exists.



# Appendix A: “AI with AI” Podcast Mindmaps

---

This appendix contains a complete set of three-level-deep time-ordered mindmaps of all research and (a selection of) news stories discussed on the “AI with AI” podcast (through August 2020), including contextually embedded hot-links to original source material.<sup>284</sup>

Since its inaugural episode on 3 November 2017, a new podcast has been published every week, with almost no breaks. As of September 2020, a total of 147 episodes have been published, most of which are between 30-40 minutes long (the longest is close to an hour). For purposes of this study, the most significant part of this resource is the 2600-plus pages of detailed notes that accompany it.<sup>285</sup>

The podcast notes are a treasure trove of information, insofar as they contain summaries of over 400 AI/ML-related research studies (including most of the “breakthrough” and “milestone events” that have occurred since the middle of 2017) and over 125 AI-related news stories that bear directly on AI/ML technologies (30 of these are included in the mindmaps that follow). Because the font-size of individual entries is typically very small, in order to accommodate their sheer volume, these mindmaps are best viewed interactively using an Adobe PDF version of this paper (rather than seeing a print copy). Viewing them in PDF form has the added virtue of having ready access to over 800 active links, such that, when clicked on with the mouse, will automatically open up an associated web-based reference.

Figure 22 shows the 22 combinations of *shape*, *boundary type* and *color*, *fill color*, and *text-color* used to highlight different categories. For example, research-oriented entries that involve SOTA advances are highlighted as SOTA Research research that highlights challenges to AI performance and/or identifies “AI failures” are depicted as Challenges; and reports, and/or policy announcements made by the US Defense Department are indicated as DOD Report.

Figure 23 summarizes yearly counts of coded entries: a total of 25 entries are from 2017, 128 entries from 2018, 177 from 2019, and 100 from January 2020 through August 2020. Tallies for individual code combinations are split into two types: (1) *Main*, which refers to entries whose *top level* description matches a given coding (i.e., that entry’s “main” or dominant discussion is about the topic implied by its coding), and (2) *All*, which includes all entries, regardless of the “level” on which they appear in the mindmap (i.e.,  $\#All \geq \#Main$ ).

---

<sup>284</sup> *AI with AI*, CNA podcast, <https://www.cna.org/CAAI/audio-video>.

<sup>285</sup> The “AI with AI” notes are compiled by the author of this white paper, who also co-hosts the podcast itself.

Figure 22. List of shape and color codes used to highlight mindmap entries

<b>State-of-the-Art (SOTA)</b>	Research that includes exceeding state-of-the-art metric(s) on a problem
<b>Milestone Achievement</b>	Research that achieves a significant new milestone on a well-studied problem
<b>Interpretability   Explainability</b>	Research that involves aspects of interpretability and/or explainability
<b>Innovative Concept   Fundamental Insight</b>	Research that introduces a new concept or makes a fundamental insight
<b>Innovative Achievement</b>	Research that achieves a notable and/or novel "one off" achievement
<b>General comments</b>	Summary points and discussion of research goals, assumptions and accomplishments
<b>Academic Announcement   Policy   Framework   Report</b>	Reports, surveys, and/or policy announcements made by <i>academic institutions</i>
<b>International Announcement   Policy   Framework   Report</b>	Reports, surveys, and/or policy announcements made by <i>international organizations</i>
<b>Industry Announcement   Policy   Framework   Report</b>	Reports, surveys, and/or policy announcements made by the <i>commercial industry</i>
<b>US Announcement   Policy   Framework   Report</b>	Reports, surveys, and/or policy announcements made by the <i>U.S. government</i>
<b>DoD</b>	Reports, surveys, and/or policy announcements made by the <i>U.S. defense department</i>
<b>SOTA Review   Meta-Analysis   Survey</b>	Technical reviews, surveys of state-of-the-art methods, and/or meta-research analysis
<b>Concept Paper   Method</b>	Technical concept papers introducing an idea or an "idea for a new method"
<b>Critique</b>	Technical critique of existing research, method, or study
<b>Hardware   Technology</b>	Research that emphasizes hardware rather than software (or new/refined ML method)
<b>AI   Neuroscience   Biology   Robotics</b>	Research that sits at the cusp between AI and neuroscience, biology, and robotics
<b>Human - AI - Neuro Interaction</b>	Research that involves direct human ↔ AI interaction (e.g., brain-computer interfaces)
<b>Swarms   Technology   Algorithms</b>	Research that focuses on swarm technology, either in software and/or robotic
<b>Physical system</b>	Research that involves instantiating an ML method in a physical system
<b>AI/ML Limitation   Vulnerability</b>	Research that focuses on identifying, mitigating, and/or leveraging AI/ML limitations
<b>'AI Challenge'   'Failure'   Questionable Gains</b>	Research that highlights challenges to AI performance and/or identifies AI failures
<b>'Anti AI' Backlash</b>	Studies, reports, public protests, and/or legislature that serve as "Anti-AI" backlash

Figure 23. "AI with AI" podcast corpus summary

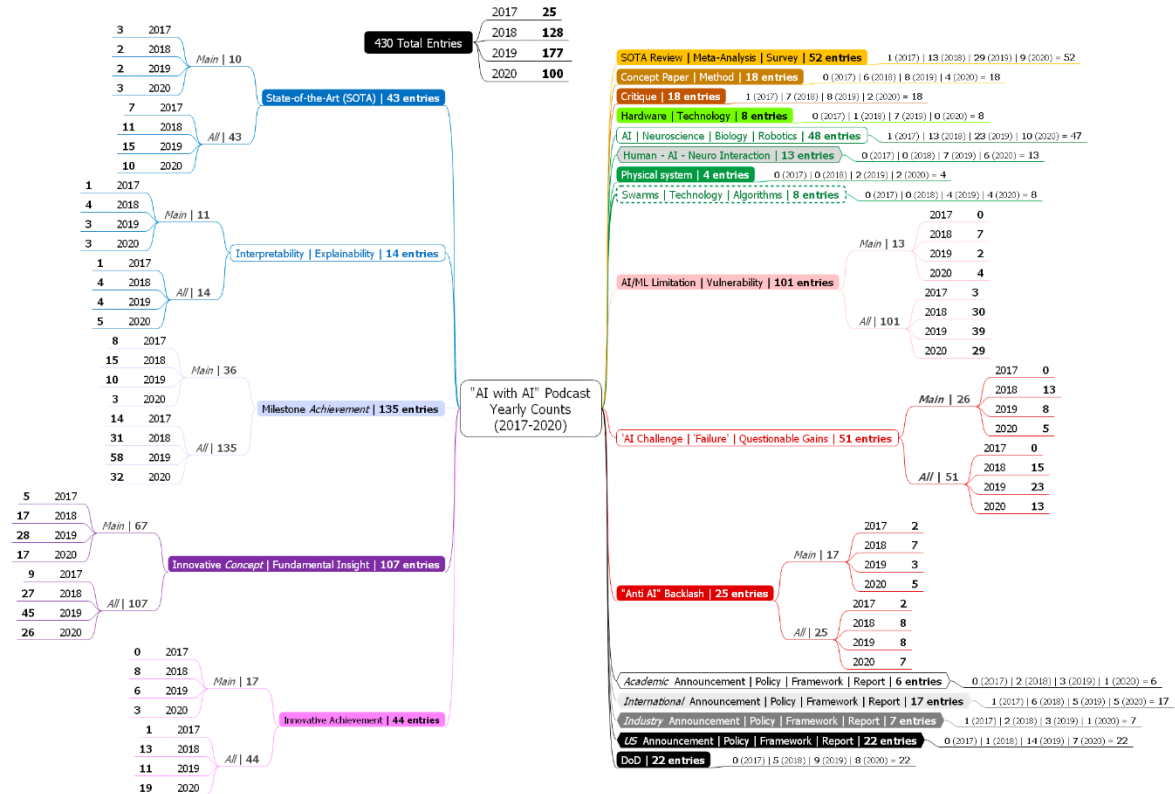


Figure 24 and Figure 25 tabulate only those entries that are coded as top-level (i.e., category “Main” in Figure 23) *SOTA*, *milestone achievements*, or *innovative concepts*. Figure 24 is a flat view of individual entries, and contains additional amplifying information about what a particular “milestone” is about (e.g., “GPT-2,” the 5<sup>th</sup> entry on the list for 2019, uses a text database, and involves text generation / NLP, is a refinement of GPT-2, and represents a SOTA advance). Figure 25 collates these amplifying comments to provide a general sense of which topics and approaches are emphasized over others (at least within the “AI with AI” corpus of studies). For the most part, the major themes represented here are consistent with the broader *ArXiv*-based survey (discussed in **Recent Trends**). The top seven domains that appear as “Main” SOTA/milestone/innovative entries include, in order: *game-centric refinements*, *combination (of existing ML methods)*, *discovery*, *pattern recognition*, *knowledge representation*, *curiosity and self-exploration*, and *general search*.

Figure 26 and Figure 27 tabulate only those entries that are coded as top-level (i.e., category “Main” in Figure 23) *challenges*, *limitations*, *vulnerabilities*, or “*Anti AI*” *backlashes*. Figure 26 is a flat view of individual entries, and (as Figure 25 did for the entries in Figure 24) Figure 27 collates these amplifying comments to provide a general sense of how “challenges” are distributed across various problem domains. The takeaway from Figure 27 is the sheer *number* of problems and problem domains for which top-level entries discuss associated challenges, limitations, and vulnerabilities; these range from fundamental challenges still facing general ML systems, to health, self-driving cars, and facial recognition (the latter of which has also seen increasingly strong pushback on an international scale).

The remaining pages of Appendix A contain high-resolution Adobe-PDF-formatted mindmaps of the entire January 2017 – August 2020 corpus of “AI with AI” podcast stories.



Figure 25. SOTA + Milestone Achievements + Innovative Concepts: Counts

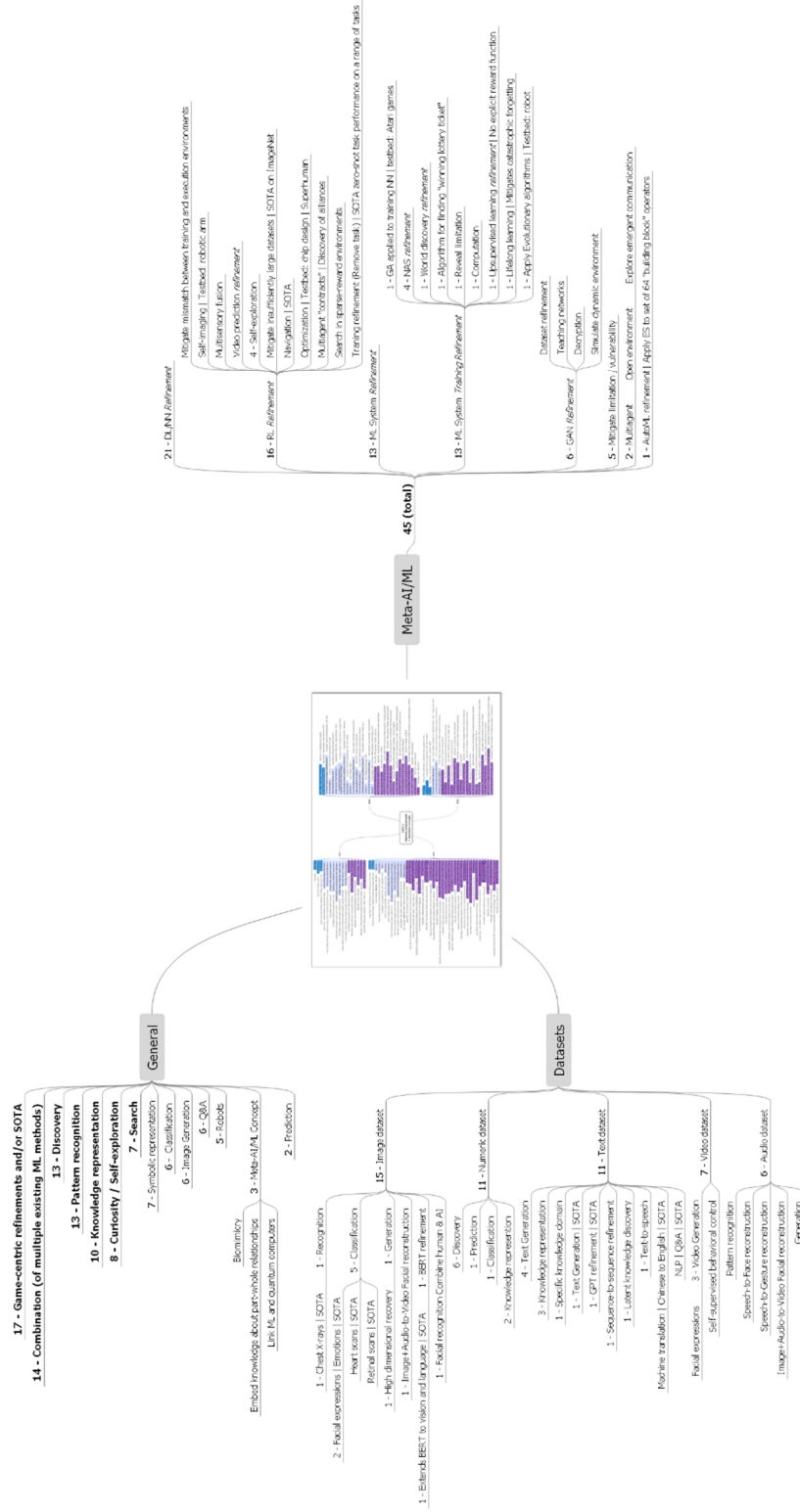
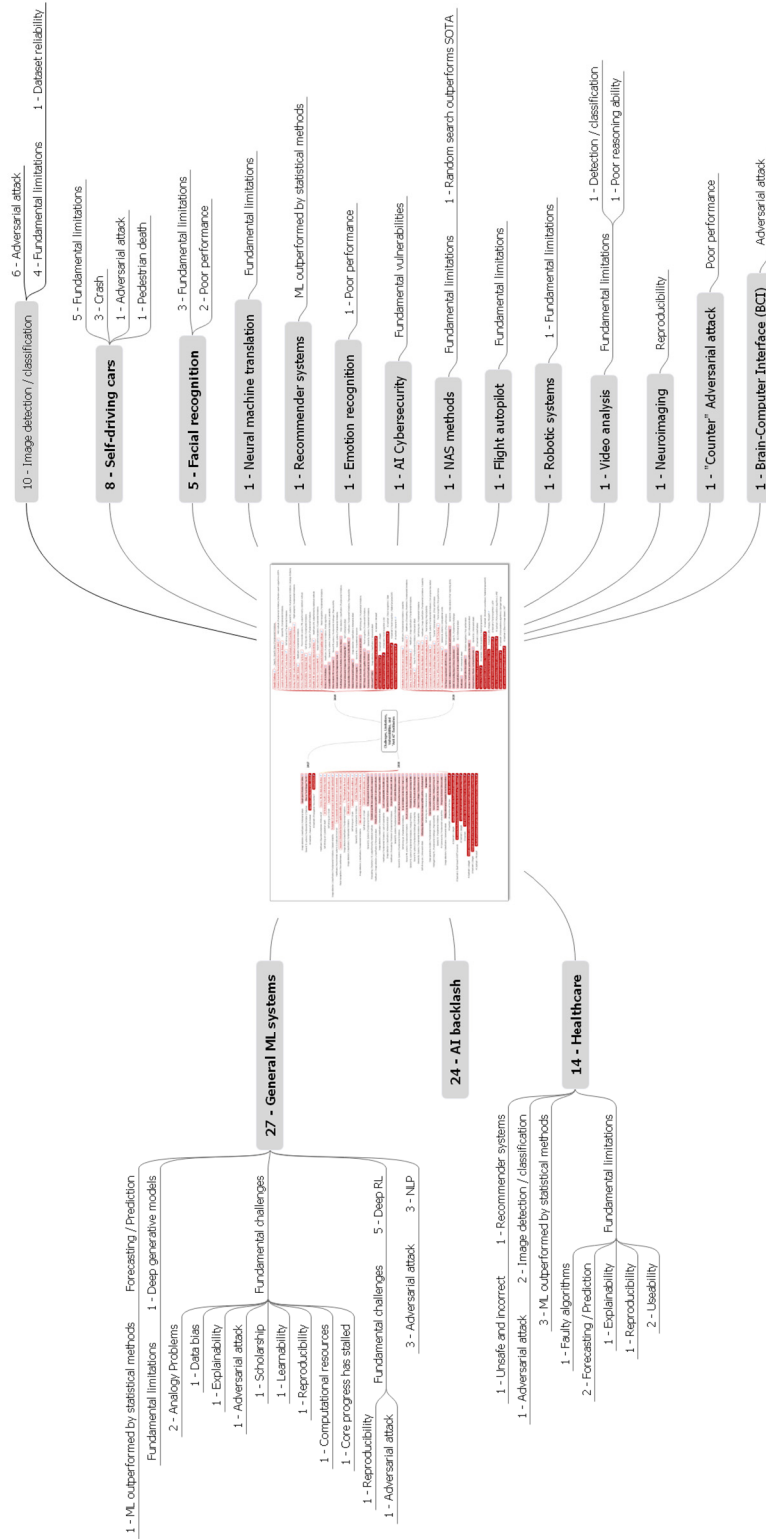


Figure 26. Limitations, Vulnerabilities, and "Anti AI" Backlashes



Figure 27. Limitations, Vulnerabilities, and "Anti AI" Backlashes: *Counts*



2017

**Project Maven** April
Launched in April 2017 by then-Deputy Defense Secretary Bob Work to accelerate the department's integration of big data, AI and ML into DoD programs
Also called the Algorithmic Warfare Cross-Function Team (AWCFT)

**AutoML** May
AutoML defeats the human AI engineers at their own game by building ML software that's more efficient and powerful than the best human-designed systems
Google

**Transformer** June
A novel NN architecture for language understanding
Outperforms both recurrent and convolutional models on academic English to German and English to French translation benchmarks
SOTA while being more parallelizable and requiring significantly less time to train
Based solely on attention mechanisms, dispensing with recurrence and convolutions
Google Brain

**Libratus defeats poker professionals** July
Difficult because poker entails imperfect knowledge of hidden information
Blueprint strategy: computes game abstraction that is small and easy to solve
Subgame solving: constructs fine-grained abstraction based on play
Real-time self-improvement: improves blueprint strategy as competition proceeds
Breaks up game into computationally manageable parts and fixes potential weaknesses in its strategy during competition
Carnegie Mellon University

**Fruit flies used to improve NN algorithms** August
Leverage biomimicry: consider how fruit flies categorize odors
Prototypes of the scheme on simple testbed problems resulted in 30%-50% accuracy improvement
They take their very high-dimensional object, and they'll try to reduce it and then look for similarity in this lower dimension of space. It's kind of like doing a principle components analysis, a popular technique. You take some data and try to plot it, for example, in two dimensions, while still preserving the structure, so that you can visualize it better. What the fly actually does is, instead of reducing it, it expands the dimension into much larger than it was, and it creates a very sparse point in a high dimensional space.
University of California, San Diego; The Salk Institute for Biological Studies, CA

**Open letter calling to ban killer robots** August
Future of Life Institute

**Survey: Agents Modelling Other Agents** September
Considers seven major methodologies for agents modelling other agents: Policy reconstruction, Type-based reasoning, Classification methods, Plan recognition, Graphical models, Recursive reasoning, Group modelling.
Identifies a number of open problems: Synergistic Combination of Modelling Methods, Policy Reconstruction under Partial Observability, Safe and Efficient Model Exploration, Efficient Discovery of Decision Factors, Computationally Efficient Implementations, Modelling Changing Behaviors, Modelling with Action Duration, Modelling in Open Multiagent Systems, Autonomous Model Contemplation and Revision.
University of Edinburgh, University of Texas at Austin

October

**AlphaGo Zero** DeepMind
Defeats AlphaGo 100-0 after 3 days of training (compared to several months for original AlphaGo) and without any human intervention/human-game-playing-data

**ML Replicates Chaotic Attractors**
Able to predict system to eight "Lyapunov times" into future, eight times further ahead than previous methods allowed!
One doesn't need equations, only data
Uses reservoir computer- introduced 15 years ago!
Ed Ott, University of Maryland

**AI solves "noisy cocktail party" problem**
Picks a single voice out of a crowd ("noisy cocktail party" problem): AI is able to separate the voices of multiple speakers in a noisy environment.
Mitsubishi Electric Research Laboratory, Cambridge, Massachusetts

**DeepXplore**
Method for error-checking reasoning of thousands or millions of neurons in unsupervised DNN
Uses network itself to generate test images likely to cause neuron clusters to make conflicting decisions
"You can think of our testing process as reverse-engineering the learning process to understand its logic"
Columbia University; Lehigh University

**Capsule Networks**
Extend the sharing of knowledge across locations to include knowledge about the part-whole relationships that characterize a given shape
Trained on a testbed of images that stressed "shape recognition"
Beat SOTA CNNs by 45%
Showed marked resistance to adversarial attack
Geoffrey Hinton

**Single-pixel Attacks Fool CNNs**
Show that over 2/3 of the natural images in Kaggle CIFAR-10 test dataset and over 16% of the ImageNet test images can be perturbed to at least one target class by modifying just one pixel (with 74% and 23% confidence on average)
Show the same vulnerability on the original CIFAR-10 dataset
Kyushu University, Japan

**Bias as "real danger" for AI**
"The real safety question, if you want to call it that, is that if we give these systems biased data, they will be biased"
John Giannandrea, AI lead at Google

November

**Robot Passes a Medical Licensing Exam**
Scored 456 points, which is 96 points above required mark
Chinese AI-powered robot Xiaoyi

**CheXNet diagnoses pneumonia better than radiologists**
Diagnoses 14 medical conditions based on chest X-ray images
Exceeded the accuracy of all previous algorithms
Stanford University

**AutoML "designs" NASNet for computer vision**
Problem: to recognize objects - people, cars, traffic lights, etc. — in a video in realtime
NASNet was 82.7% accurate at predicting images on ImageNet's validation set
1.2% better than SOTA results using human-engineered NN-based systems
Google

**1st meeting of the Convention on Conventional Weapons (CCW) Group of Governmental Experts on lethal autonomous weapons systems**
Not focused on a ban on fully autonomous weapons. According to the UN release, the meeting aims at a conversation about the legal and ethical challenges with forthcoming military technology
United Nations

**Slaughterbots**
Hard-hitting video which shows killer drones capable of zeroing in on their particular target, using facial recognition, and carrying out a coordinated mass killing in a classroom
"The Campaign to Stop Killer Robots," Future of Life Institute

December

**AlphaZero** DeepMind
Given only basic rules, but no other strategies or tactics
Defeated world champion chess-playing program, Stockfish, after 4 hours training
Required 8 hours to defeat original AlphaGo

**ML used to 'discover' exoplanets**
NASA
Demonstrates new ways of analyzing Kepler data

**ML 'predicts' chemical reactions**
Trained millions of chemical reactions to "understand" the basic structure of the 'language' of organic chemistry
Achieves accuracies up to 80% (but only for systems in which largest molecules have no more than 150 atoms)
IBM

**Deep Neuroevolution**
GA applied to training NN
Research group "surprised" to learn that "simple" GA can be used to train deep convolutional networks with over 4 million parameters to play Atari games from pixels
Required between 720 and 3000 CPUs distributed across a large, high-performance computing cluster
Uber Engineering

**Ethically Aligned Design (EAD)**
Interim version (v2) of final report due in 2019 seeks input from researchers and public at large
IEEE

**Neural networks can 'read minds'**
Generalizes and extends previous results
Study lasted 10 months and consisted of three people viewing images of three different categories: natural phenomena (such as animals or people), artificial geometric shapes, and letters of the alphabet for varying lengths of time
Model 'generates' images based on brain activity, as opposed to matching that activity to existing examples
Kyoto University



# 2018

### January

- AI "defeats" humans on reading comprehension test**
  - Stanford Question Answering Dataset (SQuAD) consists of questions posed by crowdworkers on a set of Wikipedia articles
  - Microsoft (Jan 3), followed by Alibaba (Jan 5)
- AI code decryption**
  - Without any prior knowledge, AI algorithms cracked two classic forms of encryption
    - Caesar cipher
      - Named after Julius Caesar. Suspecting enemies of eavesdropping, Caesar shift each letter in his messages by three places along the alphabet
      - For example, "Caesar" became "Fdhvdu."
    - Vigenère cipher
      - Similar to Caesar, but switches the amount of alphabet shifting with each letter
  - AI uses a generative adversarial network (GAN) that starts without any knowledge of ciphers or language
  - Google, University of Toronto
- Time-Contrastive Network (TCN)**
  - Robot learns from video
  - Imitation of human behavior requires a viewpoint-invariant representation that captures the relationships between end-effectors (hands or robot grippers) and the environment, object attributes, and body pose
  - Self-supervised
  - Trained TCN produces input to RL algorithm to control the robot's movements
  - Google, University of Southern California

### January

- Critical appraisal of deep learning**
  - 11 challenges faced by current DL systems
    - DL thus far is data hungry
    - DL thus far is shallow and has limited capacity for transfer
    - DL thus far has no natural way to deal with hierarchical structure
    - DL thus far has struggled with open-ended inference
    - DL thus far is not sufficiently transparent
    - DL thus far has not been well integrated with prior knowledge
    - DL thus far cannot inherently distinguish causation from correlation
    - DL presumes a largely static universe
    - DL thus far works well as an approximation, but its answers often cannot be fully trusted
    - Deep learning thus far is difficult to engineer with
    - Excess hype
  - Gary Marcus
- Neuromorphic computing**
  - Superconducting synapse
  - Goal is to design computer chips that work like the human brain
  - Fires 200 million times faster than human brain, uses one ten-thousandth as much energy
  - NIST, IARPA

### February

- AutoML 'Discovers' NN Architectures**
  - Achieve SOTA results for CIFAR-10, mobile-size ImageNet and ImageNet
  - Ask whether a combination of hand-design and evolution could do better than either approach alone
  - DeepMind
- Pointing and Justification Explanation (PJ-X)**
  - Feature visualization combined with other interpretability techniques helps understand aspects of how networks make decisions as a whole
  - UYC Berkeley, University of Amsterdam, Facebook AI Research
- Hindsight Experience Replay (HER)**
  - So why not just pretend that you wanted to achieve this goal to begin with, instead of the one that you set out to achieve originally?
  - HER formalizes what humans do intuitively: you have not succeeded at a specific goal, you have at least achieved a different one
  - OpenAI
- Scheduled Auxiliary Control (SAC-X)**
  - Predicated on idea that to learn complex tasks from scratch, agents must learn to explore and master a set of basic skills first
  - Researchers don't tell robot how to complete task, they simply equip it with sensors (initially turned off) and let it fumble around until it gets things right
  - DeepMind
- "Diversity is All Your Need" (DIAYN)**
  - New approach to unsupervised learning does not require an explicit reward function
  - Maximize individual utility and collective diversity
  - Learns skills by using maximum entropy policy
  - Google Brain, University at Berkeley

### February

- Evolutionary algorithm 'finds' previously unknown game hack**
  - Applied to arcade classic Q\*bert, in which players must navigate a strange orange character around a pyramid and dodge enemies
  - AI "discovers" the rule: "If you can't win, kill yourself or cheat."
  - Lesson: On the one hand, it shows how evolutionary approaches let AI succeed without human help. On the other hand, it reminds us that we may need to place limits on which strategies AIs are allowed to use in order to achieve their goals
  - University of Freiburg, Germany
- Review: ML for clinical predictions**
  - Compare deep learning techniques and LR models in prediction of health-related outcomes in traumatic patients
  - "The results ... showed that ANN has better performance than LR in predicting the terminal outcomes of traumatic patients in both the AUC and accuracy rate"
  - Questionable conclusion: judging by published confidence intervals, AUC statistics cannot be used to distinguish between ANNs and LR
  - International Journal of the Care of the Injured
- Leveraging fish-school energy dynamics to optimize autonomous swarming drones**
  - Efficient collective swimming by harnessing vortices via deep RL
  - Example of biomimicry-motivated research
  - ETH Zurich

## March

- AI matches human performance translating news from Chinese to English**
  - First machine translation system that can translate sentences of news articles from Chinese to English with the same quality and accuracy as a person
  - Parity achieved on a commonly used test set of news stories, called *newstest2017*
  - Microsoft
- DCNN outperforms human cardiologists on heart scans**
  - The AI only performed the *first step* in the analysis of a heart image and the making of a diagnosis
  - Lawrence Berkeley National Lab
- Building Blocks of Interpretability**
  - First system capable of providing natural language justifications of decisions as well as pointing to the evidence in an image
  - Method for generating multimodal explanations of output; the system can "explain itself"
  - For a given question and image, PJ-X predicts answer and multimodal explanations that both point to visual evidence for a decision and provide textual justifications
  - Google
- "Surprising Creativity" of Digital Evolution**
  - Presents substantial evidence that the existence and importance of evolutionary surprises extend beyond the natural world, and may indeed be a universal property of all complex evolving systems
  - Crowd-sourced paper by 26 artificial life and evolutionary computation researchers
- "Theory of Mind" (ToM-Net)**
  - AI taught to understand thought-process behind decisions of others like a human
  - Predicts what other AIs will do in a virtual setting
  - Critique by Shimon Whiteson (Oxford University)
  - "There is no theory, the experiments are in toy domains, and the algorithmic contribution is negligible: just some engineering of a network architecture."
  - DeepMind
- The brain may learn completely differently from what we've assumed since 20th century**
  - Identify new dendritic learning process
  - Show that learning is actually done by several dendrites, similar to the slow learning mechanism currently attributed to the synapses
  - Also show that weak synapses, previously assumed to be insignificant, play an important role in the dynamics of our brain
  - Hebb's theory has been so deeply rooted in the scientific world for 70 years that no one has ever proposed such a different approach
  - Bar-Ilan University, Israel
- 1st Pedestrian Death w/Self-driving Vehicle**
  - An Uber self-driving car killed a 49yo woman, Elaine Herzberg, in Tempe, Arizona
  - Uber
- Statistical and ML forecasting methods compared**
  - Evaluates performance on large subset of monthly time series used in M3 competition
  - ML methods are not a panacea that automatically improve forecasting accuracy
  - Easily generate implausible solutions, leading to exaggerated claims
  - The six most accurate methods are basic statistical methods, not ML
  - PLOSOne

## April

- Chemical syntheses with DNN**
  - An AI system finds correct sequence of steps to synthesize complex organic molecules
  - A task much more complex than the game of Go
  - Combine three different neural networks together with Monte Carlo Tree Search (MCTS) to perform chemical synthesis planning
  - The neural networks were trained on essentially all reactions published in the history of organic chemistry
  - University of Münster
- Anticipating Temporal Occurrences of Activities**
  - Goal is to accurately anticipate human behavior *five minutes into the future*
  - Accuracy is over 40% for short forecasting periods
  - Accuracy declines the more the algorithm needed to look into the future
  - Correctly predicts action and duration 3 minutes in future about 15% of the time
  - University of Bonn, Germany
- ML 'solves' cocktail party problem**
  - Trained a multi-stream convolutional neural network-based model to split synthetic cocktail mixture into separate audio streams for each speaker
  - Google
- Learning unsupervised learning rules**
  - Meta-learn an unsupervised representation learning update rule
  - Performance that matches or exceeds existing unsupervised learning on held out tasks
  - Update rule can train models of varying widths, depths, and activation functions
  - Demonstrate meta-learning for learning complex optimization tasks where no objective is explicitly defined
  - Google Brain
- CCW GGE on LAWS**
  - 26 countries are now calling for a ban on fully autonomous weapons
- 25 European countries sign declaration of AI cooperation**
  - Austria, Belgium, Bulgaria, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, UK, Norway
- "For a Meaningful AI" Report**
  - An Economic Policy Based on Data; Building a Network of Interdisciplinary Institutions for AI; Anticipating and Controlling the Impacts on Jobs and Employment; Ethics and AI
  - French Parliament
- AI in the UK: ready, willing and able?**
  - Calls on UK to take leading role in development of AI ethics & regulation to prevent similar scandals to Cambridge Analytica from emerging
  - UK Parliament Report on AI
- SecDef Mattis announces plans for new joint program office focused on AI**
  - "We're looking at a joint office where we would concentrate all of DOD's efforts since we have a number of AI efforts underway right now, we're looking at pulling them all together"
- Medical DL systems susceptible to adversarial attacks**
  - Inputs may be *engineered* to cause misclassification on medicine
  - Harvard
- "Poison Frogs" Attacks on NNs**
  - Identifies new vulnerability: targeted clean-label poisoning
  - In "clean label" attacks, the poison image looks totally innocuous, and is labelled properly according to a human observer. This makes it possible to poison a machine learning dataset without having any inside access to the dataset creating process
  - University of Maryland, Cornell
- Google workers urge CEO to pull out of maven work**
  - Thousands of Google employees, including dozens of senior engineers, have signed letter - with more than 3,100 signatures - protesting Google's involvement in Maven)
- AI researchers boycott Korean university over its work on "killer robots"**
  - More than 50 leading AI & robotics researchers boycott South Korea's KAIST University over the institute's plans to help develop AI-powered weapons.

2018

**Lack of explainability in health care**  
 AI able to predict and diagnose diseases, from cardiovascular illnesses to cancer, and predict related things such as the likelihood of death, the length of hospital stay, and the chance of hospital readmission  
*University of Chicago, Google, University of CA, SF, Stanford University*  
 But: predictions based on patterns in data that researchers cannot fully explain

**Autodidactic Iteration (ADI)**  
 Applies RL to solve Rubik's cube  
 University of Wyoming  
 Solves 100% of randomly scrambled cubes while achieving median solve length of 30 moves—on par w/solvers that employ human domain knowledge

**"Centaur" concept applied to facial recognition**  
 DL models achieve high accuracy  
 Intel; University of Illinois Urbana-Champaign  
 30-day unplanned readmission, In-hospital mortality, All of a patient's final discharge diagnoses, Prolonged length of stay, Chances a patient will soon die

**New general form of adversarial attacks**  
 Instead of perturbing existing data points, adversarial examples synthesized entirely from scratch  
 Microsoft, Stanford University  
 Undermines existing defenses, which are designed solely for perturbation-based attacks  
 Moreover, generative adversarial examples are able to transfer to other classifiers trained using the same dataset, and can be better at evading human detection; success rates range from 22% to 37%

**AI learns to "see in the dark"**  
 CNN significantly improves low-ISO quality while shooting at faster shutter speeds  
 Intel; University of Illinois Urbana-Champaign

**AI predicts when patients will die**  
 Demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization  
 Models outperformed traditional, clinically-used predictive models in all testbed cases  
 The potential downside of this technology is that an AI-based system can be manipulated to provide more "ethical" medical care  
 Google; Univ of Chicago Medicine; Univ of CA in San Francisco; Stanford University

**ASND (RD&A) Geurts disbands unmanned systems office**  
 Goal is to nudge unmanned systems into becoming part of everything the Navy does  
 Navy wants unmanned systems to be ubiquitous in future warfare

**AGI safety literature review**  
 Overview of challenges and design ideas considered by the AGI safety community  
 Australian National University

**AI researchers allege that machine learning is alchemy**  
 "Without deep understanding of the basic tools needed to build and train new algorithms, researchers creating AIs resort to hearsay, like medieval alchemists"  
 Science/AAAS

**Is an AI Winter On Its Way? - Part 1/2**  
 Scathing review; focuses mostly on autonomous driving systems  
 Filip Plekiewicz, Principal AI Scientist at Koh Young Technology, Inc.

**Grid-like representations in artificial agents**  
 Agents tasked with/learning spatial search reproduced "grid cells" very similar to those found in mammalian brains, despite not being programmed to appear  
 Used recurrent neural networks (RNN) with long short-term memory (LSTM)  
 Artificial agents tasked with/learning spatial search reproduced "grid cells"  
 DeepMind, University College London, UK

**Reward-based learning**  
 Investigates role the neurotransmitter Dopamine plays in learning  
 "Dopamine [is believed to] strengthen synaptic links in the prefrontal system, reinforcing particular behaviors"  
 "Shows that dopamine-like reward isn't only used to adjust weights, but it also conveys and encodes important information about abstract task and rule structure, allowing faster adaptation to new tasks."  
 "In AI, this means the dopamine-like reward signal adjusts the artificial synaptic weights in a neural network as it learns the right way to solve a task"  
 DeepMind

**Spiking Neural Network Architecture (SpiNNaker)**  
 Neuromorphic computer with 1 million processors switched on for first time  
 Capable of completing more than 200 million million actions per second, with each of its chips having 100 million moving parts  
 Institute of Neuroscience and Medicine, Jülich Research Centre, Germany

**Proof reveals fundamental limits of "knowability"**  
 Wolpert effectively proves that there is always something that the inference device cannot predict, and something that it cannot remember, and something it cannot observe  
 David Wolpert, Sante Fe Institute

**Autopilot slams into police car**  
 A Tesla car operating in "autopilot" mode crashed into stationary police car in Laguna Beach, CA, leaving driver injured and patrol car totalled  
 3rd time a Tesla in autopilot has crashed into a stationary emergency vehicle in 2018  
 Tesla

**ClariNet = NN for text-to-speech (TTS)**  
 Significantly outperforms previous separate TTS models  
 Directly converts text to a speech waveform in a single neural network  
 Baidu

**AL exceeds animals at predicting chemical toxicity**  
 Achieves 80-95% accuracy across 9 health hazards with no constraints on compounds  
 John Hopkins University

**Training ImageNet in Four Minutes**  
 Previous record time for training ResNet-50 was 15 minutes  
 Previous record for AlexNet had been 11 minutes  
 Trained ResNet-50 in 6.6 minutes; AlexNet in 4 minutes on the ImageNet dataset  
 Chinese Tencent Machine Learning

**AI perceives human emotions**  
 Outperforms traditional systems in capturing these small facial expression variations  
 MIT Media Lab

**Discovering physical concepts with NNs**  
 Combines supervised and unsupervised ML techniques with techniques developed under the broad rubric of representation learning  
 Architecture that models the physical reasoning process  
 Institute for Theoretical Physics, Zurich, Switzerland

**Measuring abstract reasoning in NNs**  
 SOTA NNs cannot "understand" problems they not explicitly trained to solve  
 Testbed: Raven's Progressive Matrices (RPM) = introduced in 1936 by the psychologist John Raven) features several rows of images with the final row missing its final image  
 DeepMind

**Troubling Trends in ML Scholarship**  
 Failure to identify the sources of empirical gains  
 Failure to distinguish between explanation and speculation  
 Misuse of language  
 Use of mathematics that obfuscates or impresses rather than clarifies  
 Carnegie Mellon University, Stanford University

**AI recommends unsafe/incorrect cancer treatments**  
 Internal company documents from IBM show that medical experts working with the company's Watson supercomputer found "multiple examples of unsafe and incorrect treatment recommendations" when using the software  
 IBM/Watson

**Amazon's Rekognition facial recognition identifies 28 lawmakers as crime suspects**  
 The test misidentified people of color at a high rate (39%) even though they made up only 20 percent of Congress  
 ACLU

**Tech leaders sign pledge promising to not develop LAW**  
 Signatories include SpaceX and Tesla CEO Elon Musk; co-founders of Google's DeepMind subsidiary, Shane Legg, Mustafa Suleyman, and Demis Hassabis; Skype founder Jaan Tallinn; and some of the world's most respected and prominent AI researchers, including Stuart Russell, Yoshua Bengio, and Jürgen Schmidhuber  
 Future of Life Institute

May

**Project Debater**  
 AI system engaged in the first ever live, public debates with humans  
 IBM  
 In both debates, audience voted Project Debater to be worse at delivery but better in terms of the amount of information it conveyed  
 Created to debate nearly 100 topics (but technology can carry out a meaningful debate only about 40% of the time); the debate topics were chosen by IBM!

**OpenAI Five defeats humans in Dota 2**  
 Consists of five single-layer, 1,024-unit long short-term memory (LSTM) networks  
 OpenAI  
 Dota 2 is a popular multiplayer online battle arena strategy game  
 To prep for the matches, the system plays 180 year's worth of games every day  
 It handily beat the first three teams in several rounds, and it won two of the first three games against the fourth and fifth squads  
 OpenAI Five developed strategies mirroring those of professional players

**Deep Curiosity Search (DCS)**  
 Encourages intra-life exploration by rewarding agents for visiting as many different states as possible within each episode  
 Uber AI Labs  
 Matches performance of current SOTA methods on Montezuma's Revenge (MR)

**AutoAugment**  
 Problem: collecting enough quality data to train model well often too difficult  
 Proposal: use RL to increase amount / diversity of data in existing training dataset  
 Heuristic: images have many symmetries that don't change information content  
 Method improves ImageNet to new SOTA accuracy  
 Google

**Atom2Vec: recreates periodic table of elements in hours**  
 Atom2Vec: a form of unsupervised natural language processing  
 Atom2Vec learned to distinguish between different atoms after analyzing a list of chemical compound names and then used natural language processing to cluster the elements according to their chemical properties  
 Stanford

**New conceptual framework for ML**  
 Combinatorial generalization top priority for AI to achieve human-like abilities  
 Structured representations and computations are key to realizing this objective  
 Present new building block for AI toolkit: the graph network (GN)  
 Main unit of computation in the GN framework is the GN block a "graph-to-graph" module which takes a graph as input, performs computations over the structure, and returns a graph as output  
 27 authors; DeepMind, Google Brain, MIT, University of Edinburgh

**DoD requests \$70M to establish JAIC**

**New fundamental rule of brain plasticity**  
 When one connection, a synapse, strengthens, immediately neighboring synapses weaken based on the action of a crucial protein called Arc  
 Explains how synaptic strengthening and weakening combine in neurons to produce plasticity  
 Picower Institute for Learning and Memory at MIT

**Reliability of current measures of progress in ML questioned**  
 Replication study: how well do classifiers generalize to new unseen data from same distribution, w/CIFAR-10 testbed (subject of intense research for about 10 years)  
 Results cast doubt on the robustness of current classifiers  
 Models learn datasets and datasets only, not the reality that generated those datasets  
 MIT, Berkeley

**Google does not renew contract for Project Maven**  
 Company won't pursue 'similar' deals to controversial Maven

June

**AI matches health experts at spotting eye diseases**  
 System that can analyze retinal scans and spot symptoms of sight threatening eye diseases  
 Uses two networks  
 Segmentation network, converts the raw OCT scan into a 3D tissue map  
 Classification network analyzes the 3D tissue map and makes decisions about what the diseases might be and how urgent they are  
 System can quickly examine optical coherence tomography (OCT) scans and make diagnoses with the same or better accuracy as human clinicians  
 DeepMind

**Variational Autoencoder with Shared Embeddings (VASE)**  
 New method for lifelong learning to mitigate catastrophic forgetting  
 Based on the Minimum Description Length (MDL) principle  
 Formalization of Occam's razor in which the best hypothesis for a given set of data represents the optimal compression of the data  
 VASE automatically detects shifts in training data distribution and uses this information to allocate spare latent capacity to novel dataset-specific disentangled representations, while reusing previously acquired representations of latent dimensions  
 DeepMind

**Neural Arithmetic Logic Units (NALU)**  
 Represents numerical quantities as individual neurons without a nonlinearity  
 Learns functions that capture underlying numerical nature of data and generalizes to numbers several orders of magnitude larger than those observed during training  
 May provide insights into the numerical reasoning mechanisms in humans and animals  
 DeepMind, University at Oxford, University College London

**2nd Convention on Conventional Weapons (CCW) Group of Governmental Experts on LAWS**  
 United Nations

**White Paper: Mitigating risks of military AI**  
 Part I identifies how military use of AI could create unexpected dangers and risks  
 ML Systems Can Be Easily Fooled or Subverted  
 AI Systems Are Vulnerable to Hacking  
 Reinforcement Learning Systems Have Unpredictable Dynamics  
 Automation of Escalation Pathways  
 Part II offers and elaborates on an agenda for mitigating these risks  
 Create an information network amongst other communities, labs, and nations  
 Focus ML development on processes outside "kill chain" like logistics, systems diagnostics, and defensive cyber (giving defender greater control in virtual combat zone/minimize risk)  
 Conduct more intensive research on ML predictability, robustness, and safety  
 Electronic Frontier Foundation

**Unmanned Systems Integrated Roadmap 2017-2040**  
 Focus topics: Interoperability, Autonomy, Network Security, Human-Machine Collaboration

**Study: Curiosity-Driven Learning**  
 Testbed: 54 standard benchmark environments (that include exemplars of video games, physics engine simulations, and virtual 3D navigation tasks) in which AI models are trained without any "extrinsic rewards" or end-of-episode signals  
 Results show surprisingly good performance / strong alignment between intrinsic curiosity objective and hand-designed extrinsic rewards of many game environments  
 The motivation is simply to experience new things  
 OpenAI, UC, Berkeley, University of Edinburgh

**Survey | Critique: Explanation in AI**  
 Most work in explainable AI uses only researchers' intuition of a 'good' explanation  
 Surveys vast bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations  
 Discusses ways that these can be infused with work on explainable AI  
 Tim Miller, University of Melbourne, Melbourne, Australia

August

**OpenAI's bots "defeating humans" at Dota 2 questioned**  
 Main criticism: The OpenAI Five bots played Dota 2 by reading the game's information directly from its application programming interface (API), which allows other programs to easily interface with Dota 2  
 Effectively gives AI instant knowledge, while humans must visually interpret screen  
 While the API is designed not to give the AI more information than a human would have, it is still the case that what the AI is able to know, it knows perfectly and instantaneously  
 Motherboard, Twitter thread

**"The Elephant in the Room"**  
 Identify family of common failures of SOTA image detectors  
 Experiments involve "transplanting" object from one image in a new location of another image  
 Slight changes in object position can affect its identity according to an object detector as well as that of other objects in the image  
 York University, University of Toronto

**Google employees protest company's secret censored search engine (Dragonfly) for China**  
 Call for ethical assessment of Google projects including Dragonfly and Maven

2018

September

**GAN: High Fidelity Natural Images** (DeepMind)

- Set new SOTA in class-conditional image synthesis
- Demonstrate that GANs benefit dramatically from scaling
- Models become amenable to "truncation trick," a sampling technique that allows explicit, fine-grained control of the tradeoff between sample variety and fidelity
- Discover instabilities specific to large scale GANs

**Proposed new approach is to use NNs that are, due to their structure, inherently impervious to adversarial attacks, even when trained on standard input only**

**Radio-based system can estimate 2D poses through walls despite never trained on scenarios** (RF-Pose: Seeing Through Walls with Wi-Fi Signals) (MIT CSAIL)

- Operating distance dependent on transmission power

**Ethics Certification Program for Autonomous and Intelligent Systems** (IEEE)

- Central themes: accountability, transparency and reduction of algorithmic bias
- One of the first programs dedicated to the creation of an A/IS certification process

**DeepMind Safety Research** (DeepMind)

- Guidelines aim to prevent AI/ML algorithms from learning and figuring out their own solutions to problems in unexpected and unwanted ways

**Survey: Learning causality with data** (Arizona State University)

- Review traditional & frontier methods and open problems of learning causality

**Survey: Adversarial Attacks and Defences** (Ohio State University, Columbus; Indian Institute of Technology, Kharagpur)

- Detailed discussion of different types of adversarial attacks with various threat models and also elaborate the efficiency and challenges of recent countermeasures against them

**BrainNet** (University of Washington, Carnegie Mellon University)

- Allows a small group to play a collaborative Tetris-like game
- 1st "social network" of brains allows 3 people to transmit thoughts to each other

**Attacking object detectors via imperceptible patches on background** (State University of NY, Albany)

- New vulnerability of object detectors (OD): adding minimal perturbations to small background patches outside of targets to fail the detection results
- Misleads ODs by simultaneously decreasing true positives and increasing false positives
- Confirms that state-of-the-art image detection systems used in self-driving cars are vulnerable to attack

**ShapeShifter** (Georgia Tech)

- Generate perturbed stop signs that can consistently fool Faster R-CNN in real drive-by tests (videos available on the GitHub repository), calling for imperative need to improve and fortify vision-based object detectors
- ShapeShifter is the first targeted physical adversarial attack on Faster R-CNN object detectors (found in many SOTA detectors)

December

**AlphaFold: Using AI for scientific discovery** (DeepMind)

- AlphaFold - under development for past two years - builds on years of prior research in using vast genomic data to predict protein structure
- Beat 98 competitors by predicting structure of 25 / 43 proteins
- The 2nd place system could only predict three protein structures
- AlphaFold forecast its first protein structures in hours; magnitudes faster than previous systems (SOTA)
- AlphaFold did not "lap the field": it won a lot of rounds, with an average margin of 15% accuracy improvement over other groups on the toughest 43 tests
- DeepMind's 1st milestone in showing how AI research can drive and accelerate new scientific discoveries
- AlphaFold submitted to the Critical Assessment of Structure Prediction (CASP) protein-folding competition
- Though DeepMind had not previously competed, AlphaFold easily defeated competing solutions

**GAN Dissection** (MIT-IBM Watson AI Lab)

- Studies the internal representations of GANs
- Present analytic framework to visualize and understand GANs
- Examine contextual relationship between these units and their surroundings by inserting discovered object concepts into new images
- Quantify causal effect of interpretable units by measuring ability of interventions to control objects in the output
- By turning "on" and "off" various "neurons" and asking the GAN to paint what it thought, the researchers found distinct neuron clusters that had learned to represent a tree, for example

**Survey: Safety & Trustworthiness of DNNs** (University of Liverpool, Oxford)

- Trustworthiness = Certification (verification + testing) + Explanation
- Surveyed 178 papers between 2017-2018

**Survey: AI SOTA** (Community driven database)

- "Measuring the Progress of AI Research"
- Archive and provide links to tasks, datasets, metrics, and SOTA results
- Initial effort, pushed by group of graduate students from MIT and Harvard

**Survey: National AI Strategies** (CIFAR, a Canadian-based research institute)

- 18 AI strategies
- Provides framework for understanding the different types of strategies
- AI strategies are described according to eight areas of public policy where they are intended to have impact: scientific research, talent development, skills development, industrialization, ethics, data and digital infrastructure, government services, and inclusion

**Critique: DL vs. Symbolic Manipulation** (Gary Marcus, Yann LeCun, Jeff Dean, Judea Pearl)

- As key ingredients of AGI: spirited and voluminous witter debate
- Advocates avoiding regulating AI research, but moving to regulating AI applications in transportation, medicine, politics, and entertainment
- Balances benefits of research with potential harms of AI systems

**"Should AI Technology Be Regulated?"** (Communications of the ACM, December 2018, Vol. 61 No. 12)

- Offers five guidelines for regulating AI applications
- An AI is subject to the full gamut of laws that apply to its human operator
- An AI shall clearly disclose that it is not human
- AI shall not retain/disclose confidential data without explicit approval from source
- AI must not increase any bias that already exists in our systems
- Don't weaponize AI

**AI Now Report** (AI Now Institute at New York University)

- Includes ten recommendations focused on regulation and transparency
- Also calls for immediate action on facial recognition software
- Calls for accountability and oversight in the AI industry

**European Commission Releases Report on AI and Ethics** (Preliminary staff of about 30 individuals)

- The High-Level Expert Group on AI of the European Commission issues report that proposes framework for developing trustworthy AI
- JAIC will coordinate all DOD AI-related projects above \$15 million
- JAIC "Open for Business"

**Amoeba finds approximate solution to Traveling Salesman Problem** (Keio University, Japan)

- The problem is NP-hard
- Method harnesses amoeba's natural tendency to seek out a stable equilibrium
- One of several increasingly common examples of "non traditional" computing

**Morphogenesis in robot swarms** (Institute of Sci & Tech, Spain; Univ. of Bristol, UK; Univ. of Amsterdam, Netherlands)

- Researchers demo self-organizing swarms of 300 robots not following preset pattern
- Developed by Harvard University between 2010-2012
- kilobots = 3.3 cm tall low-cost swarm robots
- Comms via infrared
- Movement, robot-to-robot comms, and multicolor LED for monitoring
- Robots programmed to act like cells in tissue; i.e., self-organized patterning

**AI system mimics how humans visualize and identify objects** (UCLA, Stanford)

- System able to build detailed model of human body without external guidance and without the images being labeled
- Leverage what we know about how humans understand that they are looking at
- Drew insights from cognitive psychology and neuroscience

**Hiding adversarial examples from interpretation** (University of Maryland)

- Paper shows that it is possible to create adversarial patches that not only fool the prediction, but also change what we interpret regarding the cause of prediction

**Adversarial Approach to Uncover Catastrophic Failures** (DeepMind)

- Focus on two problems: searching for scenarios when learned agents fail and assessing their probability of failure
- New method finds catastrophic failures and estimates failures rates of agents multiple orders of magnitude faster than standard evaluation schemes, in minutes to hours rather than days
- Approach 1: use AI itself to figure out what befuddles AI
- Approach 2: train a neural network to avoid a whole range of outputs, to keep it from going entirely off the rails and making really bad predictions

**Autopilot crashed into police car** (Tesla)

- The accident took place despite reports that there were "100 meters of traffic cones and flashing warning lights" placed behind the police car
- A Tesla Model S on autopilot slammed into Police Car Despite in Hsinchu County, Taiwan

October

**Meta-Analysis: spurious samples in deep generative models** (AdaNet, Google, Courant Institute of Mathematical Sciences, New York)

- General framework for both learning a NN architecture and learning to ensemble to obtain even better models
- Ask: "Is it possible to get rid of all spurious samples without sacrificing the coverage of a model?"
- Spurious modes: model generates objects that clearly do not belong to the domain
- Missing modes (or the lack of coverage): does model generate all objects of the domain?
- One cannot eliminate spurious samples without sacrificing the model's ability to generate some data we actually want to model.
- CNRS/Universite Paris-Saclay

**Survey & Critique: Multiagent Deep RL** (Borealis AI, Edmonton, Canada)

- Open challenges:
  - Reproducibility, troubling trends and negative results
  - Implementation challenges and hyperparameter tuning
  - Computational resources
- Open questions:
  - Finding simplest context that exposes the innovative research idea remains challenging
  - Sparse and delayed rewards
  - On the role of self-play
  - Combinatorial nature of possible states and actions

**Is an AI Winter On Its Way? - Part 2/2** (Filip Plekiewicz, Principal AI Scientist at Koh Young Technology, Inc.)

- Update to a series of provocative blog/essays posted in May 2018
- Not just about how we are not at AGI, but a critique/warning about what we think existing techniques can/will give us
- Notes recent shut-down of Rethink Robotics, arguing that this shows that making robots do anything beyond what they already do very well in controlled factory production lines is not only difficult technically but also poses a challenging business case, even with the experience of Rodney Brooks

**Weapons Systems Cybersecurity** (GAO)

- Holistic/systems-focused discussion of multiple vulnerabilities
- The testers also looked for previously reported vulnerabilities: only one in 20 had been fixed

**New Theory of Intelligence** (Jeff Hawkins, Numenta, Inc.)

- New framework posits that every part of human neocortex learns complete models of objects and concepts
- Hypothesizes that grid cell-like neurons exist in every column of the human neocortex
- Proposes a new type of neuron called the displacement cell, which acts as a complement to grid cells, and is also located throughout the neocortex
- Grid cells are place-modulated neurons that enable an understanding of position
- Every cortical column learns models of complete objects by combining input with a grid cell-derived location, then integrating over movements
- An innovative theory that challenges conventional views and may impact both AI and neuroscience in the future

**NN unable to "understand" optical illusions** (University of Louisville)

- Compiled database of over 6,000 images of optical illusions and then trained NN to recognize them
- The results were disappointing
- "The only optical illusions known to humans have been created by evolution (e.g., eye patterns in butterfly wings) or by human artists"
- In both cases, humans provide valuable feedback; they can see illusion, but machine-vision systems cannot

**Autopilot crashed into stalled car** (Tesla)

- Car was moving at about 80 miles per hour on a Florida freeway
- Tesla: "When using Autopilot, it is the driver's responsibility to remain attentive to their surroundings and in control of the vehicle at all times"

**Google drops out of DoD's Joint Enterprise Defense Infrastructure (JEDI) competition**

- Google said in a statement: "We couldn't be assured that [the JEDI deal] would align with our AI Principles and second, we determined that there were portions of the contract that were out of scope with our current government certifications."

**Microsoft employees post open letter to prevent company from bidding on JEDI**

- The letter accuses Microsoft executives of betraying the company's artificial intelligence principles—ones that state AI should be "fair, reliable and safe, private and secure, inclusive, transparent, and accountable—in pursuit of short-term profits."

November

**Go-Explore advances SOTA on Montezuma's Revenge and Pitfall** (Open AI)

- Go-Explore differs radically from other deep RL algorithms
- Does not require human demonstrations
- Algorithm endowed with a curiosity
- Coupled with a "remedy" to the detachment problem: wherein algorithms forget about promising areas they have visited, meaning they do not return to them to see if they lead to new states

**World's first AI news anchor unveiled in China** (China's state news agency Xinhua and search engine firm Sogou)

- An AI system synthesizes the presenters' voices, lip movements, and expressions, modeled after those of real presenters

**AI Physicist for Unsupervised Learning** (Tailin Wu and Max Tegmark, MIT)

- Introduces "AI Physicist" - an unsupervised learning agent - centered on learning and manipulation of theories, which parsimoniously predict both aspects of the future (from past observations) and domain in which these predictions are accurate
- Derives relevant laws for 90% of all 40 "mystery worlds" used in testbed
- Investigates opportunities and challenges for improving unsupervised ML using four common strategies with a long history in physics: divide-and-conquer, Occam's Razor, unification and lifelong learning

**Rethinking ImageNet Pre-training** (Facebook AI Research)

- Questions paradigm of pre-training even further by exploring the opposite regime
- The prevailing paradigm is to pre-train models using large-scale data (e.g., ImageNet16) and then fine-tune the models on target tasks w/less training data
- Find that competitive object detection and instance segmentation accuracy is achievable when training on Microsoft Common Objects in Context object detection database from random initialization (from scratch without any pretraining)
- More surprisingly, achieve these results by using baseline systems and their hyper-parameters that were optimized for fine-tuning pre-trained models
- Can train large models from scratch (up to 4x larger than ResNet101) without overfitting

**Dataset Distillation** (Facebook AI Research, MIT, University of California, Berkeley)

- Compressed 60k MNIST into 10 synthetic images
- The accuracy of the model trained on solely those 10 images is above 90%
- Both practical and philosophical ramifications
- Knowledge Distillation
- Dataset pruning, core-set construction
- Understanding datasets

**BERT** (Google AI)

- Bidirectional Encoder Representations from Transformers
- SOTA on 11 NLP tasks, including the Stanford Question Answering Dataset (SQuAD v1.1)
- Open sourced

**US Considers AI Regulation for Foreign Countries** (Department of Commerce (DoC))

- Lists areas of AI that could potentially require a license to sell to certain countries
- These categories are very broad; e.g., computer vision and NLP
- Also lists military-specific products: adaptive camouflage and surveillance technology
- DOC seeks information on essentially all (i.e., broadest) AI emerging technologies

**Survey: Data Collection for ML** (School of Electrical Engineering, KAIST, Daejeon, Korea)

- Comprehensive study of data collection from a data management point of view
- Provides research landscape of operations, guidelines on which technique to use when, and identifies interesting research challenges

**Bias and generalization in GANs** (Stanford University)

- Suggests adopting experimental methods from cognitive psychology to characterize the generalization biases of machine intelligence
- Using new framework, systematically evaluates generalization patterns of SOTA models such as GAN
- Some patterns show striking similarities to experiments in cognitive psychology

**Artificial neuron implemented on quantum processor** (Universita di Pavia, Italy)

- World's first perceptron implemented on a quantum computer

**AI Hits 'Barrier of Meaning'** (Melanie Mitchell, Professor of Computer Science at Portland State University)

- ML algorithms don't yet understand things the way humans do — with sometimes disastrous consequences
- Programs that "read" documents and answer questions about them can be fooled into giving wrong answers when short, irrelevant snippets of text are appended
- Programs that recognize faces and objects can fail dramatically when their input is modified even in modest ways by certain types of lighting, image filtering and other alterations that do not affect humans' recognition abilities in the slightest.
- Programs that have learned to play particular video or board game at "superhuman" level are completely lost when the game they have learned is slightly modified

**NNs easily fooled by strange poses of familiar objects** (Adobe and Auburn University)

- Objects easily recognized by DNNs in canonical poses incorrectly classified for 97% of pose space



**AI achieves human-level performance on 3D CTF game**

DeepMind

Quake III agents individually played over 450,000 games, roughly the equivalent of roughly four years of experience

The only time humans beat a pair of bots was when they were part of human-bot team, and even then, they typically won only five percent of their matches

Used a tournament-style evaluation to demonstrate that an agent can achieve human-level performance in a three-dimensional multiplayer first-person video game, Quake III Arena in Capture the Flag mode, using only pixels and game points scored as input

**Neuro-Symbolic Concept Learner (NSCL)**

Hybrid Connectionist/Symbolic Approach

Learns by simply looking at images and reading paired questions and answers

Add more symbolic structure, and feed the neural networks a representation of the world that's divided into objects and properties rather than feeding it raw images

NSCL learns visual concepts, words, and semantic parsing of sentences without explicit supervision on any of them

**"Lottery Ticket" Hypothesis (LTH)**

MIT, MIT-IBM Watson AI Lab, and DeepMind

LTH = dense, randomly-initialized, feed-forward networks contain subnetworks ("winning tickets") that - when trained in isolation - reach test accuracy comparable to original network in similar number of iterations

"We consistently find winning tickets that are less than 10-20% of the size of several fully-connected and convolutional feed-forward architectures for MNIST and CIFAR10"

**Exploratory Look-Around**

Facebook AI; University of Texas, Austin; University of CA, Berkeley

An agent must actively observe a small fraction of its environment so that it can predict the pixelwise appearances of unseen portions of the environment

Propose RL method for which an agent is rewarded for reducing uncertainty about the unobserved portions of its environment

**Adversarial examples not bugs: they are features**

Facebook AI; University of Texas, Austin; University of CA, Berkeley

Provides evidence that adversarial vulnerability is caused by non-robust features and is not inherently tied to the standard training framework

**Predicting Future Person Activities & Locations in Videos**

Carnegie Mellon University, Google AI, and Stanford University

Propose a multi-task learning model called *NexT* the intention in terms of a predefined set of 30 activities

Propose an end-to-end, multi-task learning system utilizing rich visual features about human behavioral information and interaction with their surroundings

**Speech2Face**

MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL)

Method to reconstruct some people's rough likeness based on short audio clips

In the initial release, the software able to prove 58% of the training set

Trained on a set of 10,200 theorems

Present HOLIST: an environment, benchmark, and DL driven automated theorem prover for higher-order logic

**DeepHOL: Automated Theorem prover**

Google Research

**AI curriculum for government decision-makers**

Microsoft

Do good; for humanity; be responsible; control risks; be ethical; be diverse; open and share

Use wisely and properly; informed-consent; education and training

R&D

Use

Governance

Optimizing employment; harmony and cooperation; adaptation and moderation; long-term planning

Calls for international cooperation

**China releases code of AI ethics**

42 countries agree to AI principles

Organisation for Economic Co-operation and Development (OECD) unveiled the first intergovernmental standard for AI policies

OECD's 36 member countries including America have initially signed on along with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania

**NIST releases RFI on Development of technical standards for AI**

Seeks to understand current and potential future role of Federal agencies regarding the existence, availability, use, and development of AI technical standards and tools in order to meet the nation's needs

Seeks to understand needs and challenges regarding the existence, availability, use, and development of AI standards and tools

**The "AI in Government Act" (AIG)**

Bill charges the General Services Administration with establishing a Center of Excellence to provide expertise, as well as "conduct forward-looking, original research on federal AI policy and promote U.S. competitiveness"

Also requires agencies to establish governance plans for using AI while protecting "civil liberties, privacy and civil rights" and creates advisory board to weigh in on challenges and opportunities for agencies in adopting AI

**SF Bans FR Software**

The SF Board of Supervisors voted 8-1 on 14 May to approve the Stop Secretive Surveillance ordinance, which outlaws the use of facial recognition software or retention of information obtained through facial recognition software systems

A 2016 study by Georgetown University found about 64 million Americans were in at least one database in "a virtual line-up" even if most had no criminal record, and that there was little knowledge of whether these systems were accurate or unfairly impacted minorities

In theory, if you could get rid of all the spurious correlations in a ML model, you would be left with only the "invariant" ones—those that hold true regardless of context

Proposes how existing AI techniques can analyze causal relationships in data

**Learning Representations using Causal Invariance**

Leon Bottou, Facebook AI

Meta-learning the learning algorithms themselves

Generating effective learning environments

Meta-learning architectures

AI-generating algorithms, an alternate paradigm for producing AgI

**AI-generating algorithms**

Jeff Clune, Uber AI Labs, University of Wyoming

May have major ramifications for R&D development of novel ML algorithms

Optimal play in real-world Magic is at least as hard as the Halting Problem, solving a problem that has been open for a decade

**"Magic: The Gathering" is Turing Complete**

Georgia Institute of Technology; University of Pennsylvania

Vision + Language system + ...

A PFC (mimicking human Prefrontal Cortex, implemented by an LSTM) that combines inputs of both language and vision representations, and predicts text symbols and manipulated images accordingly

Propose an LGI network to incrementally learn meaning and usage of numerous words and syntaxes, aiming to form a human-like machine thinking process

**Language guided imagination (LGI)**

Oxford University

Quite similar to the cumulative learning process of the human being

Successfully acquired eight different syntaxes (i.e., 8 different tasks)

The degree to which neurons preferred certain numbers nearly identical to previous data from the neurons of monkeys

AI was correct 81% of the time, performing about as well as humans and monkeys do on similar matching tasks

A DNN trained for visual object recognition spontaneously developed *number sense* - innate ability of humans and other animals to assess number of #visual items in a set

**Spontaneous emergence of "number sense"**

University of Tübingen, Germany

**California legislature bars FR for police body camera**

Passed (vote: 42-18) a three year ban on state and local law enforcement from using body cameras with facial recognition

**AF Releases 2019 AI Strategy**

Delivering AI-enabled capabilities that address key missions

Partnering with leading private sector technology companies, academia, and global allies

Four Strategic Focus Areas

- Cultivating a leading AI workforce
- Leading in military ethics and AI safety
- Drive down technological barriers to entry
- Recognize and treat data as a strategic asset

Five guiding principles

- Democratize access to AI solutions
- Recruit, develop, upskill, and cultivate our workforce
- Increase transparency & cooperation w/international, government, industry, and academic partners

**DoD seeks 'ethicist' to oversee military AI**

Represents JAIC's first publicly acknowledged foray into developing AI and machine learning tools for direct warfighting applications

JAIC's "biggest project" in FY-20

**JAIC: "AI for Maneuver and Fire"**

**An Overview of State AI Initiatives**

Of the United Nation's 193 member States, only 41 States are represented

20 States have government investments in AI

There are 19 AI national plans published with 22 additional States interested in or actively developing one up to early June 2019

**Review: Neural Recommender Systems**

FutureGrasp

Considered 18 algorithms from top-level research conferences

6/7 outperformed by simple heuristic methods (e.g., nearest-neighbor methods)

Only 7 of them could be reproduced with reasonable effort

**Advanced Vehicle Technology Study**

Polytechnic University of Milan, Italy, and University of Klagenfurt, Austria

Objective #1: conduct large-scale real-world driving data collection that includes high-definition video to fuel the development of deep learning based internal and external perception systems

Objective #2: gain a holistic understanding of how human beings interact with vehicle automation technology by integrating video data with vehicle state data, driver characteristics, mental models, and self-reported experiences with technology

Objective #3: identify how technology and other factors related to automation adoption and use can be improved in ways that save lives

**Nervana Neural Network Processor (NPP)**

Intel

The Nervana NNP-T, codenamed Spring Crest, will be used for training and comes with 24 Tensor processing clusters that have been specifically designed to power neural networks

Intel's new system on a chip (SoC) provides users with everything they'll need to train an AI system on dedicated hardware

**All-optical neural network for DL**

The Hong Kong University of Science and Technology

"Proof of concept" showed that all-optical neural network is as accurate as a well-trained computer-based neural network

Identify four functionally distinct spike-waveform-based cell classes in primate cortex

Understanding function of different neuronal cell types key to understanding brain function

**Four distinct types of brain cells identified**

MIT; University of Tübingen, Germany

Tianjic is engineered so that its individual processing units can switch from spiking communications back to binary and perform a large range of calculations, in almost all cases faster and more efficiently than a GPU can

Tianjic chip integrates computer-science-oriented and neuroscience-oriented approaches to developing AgI

**Towards AgI with hybrid Tianjic chip**

University of California, Santa Barbara; Tsinghua University, Beijing, China

Triggers are transferable across models: trigger generated for GPT-2 117M model also works (in fact, even better) for the 345M model

Demonstrate input-agnostic sequences of tokens that trigger a model to produce a specific prediction when concatenated to any input from a dataset

**Universal Adversarial Triggers for Attacking NLP**

Allen Institute for AI; University of Maryland; University of California, Irvine

May

2019

August

**Pluribus: Superhuman Poker-Playing AI**

Carnegie Mellon University; Facebook AI

Pluribus victorious after a marathon 12-day poker session during which it played five human professionals at a time

First time AI has achieved superhuman performance in a multiplayer game

The core of Pluribus's strategy computed via self-play: the AI plays against copies of itself, without any data of human or prior AI play used as input

During play against opponents, Pluribus improves upon its blueprint strategy by searching for better strategy in real time for the situations it finds itself in during the game

**Learning Individual Styles of Conversational Gestures**

MIT; University of California, Berkeley

Given audio speech input, generates plausible gestures to go along with the sound

Audio does not directly encode high-level language semantics that may allow us to predict certain types of gesture (e.g. metaphors); does not separate the speaker's speech from other sounds (e.g. audience laughter)

**European Commission launches pilot phase of AI ethics guidelines**

Issue 11 specific recommendations

**Axon bans facial recognition systems**

Axon supplies 47 out of the 69 largest police agencies in the United States

"Face recognition technology not currently reliable enough to ethically justify its use"

**National AI R&D Strategic Plan: 2019 Update**

The original strategic plan was issued by the Obama administration in 2016, with President Trump calling for an update through a February 2019 executive order

Establishes objectives for Federally funded AI research, identifying the following eight strategic priorities

- Make long-term investments in AI research
- Develop effective methods for human-AI collaboration
- Understand and address the ethical, legal, and societal implications of AI
- Ensure the safety and security of AI systems
- Develop shared public datasets and environments for AI training and testing
- Measure and evaluate AI technologies through standards and benchmarks
- Better understand the national AI R&D workforce needs
- Expand public-private partnerships to accelerate advances in AI

**House Armed Services Committee Doubles Joint AI Funding**

Though this is still less than the Pentagon's request (= \$208.8 million)

Requires a second, streamlined acquisition process specifically for software

**JAIC open sources natural disaster satellite imagery dataset**

Carnegie Mellon University's Software Engineering Institute, CrowdAI, JAIC and the Defense Innovation Unit (DIU) — shared plans to open-source a labeled data set of some of the largest natural disasters in the past decade

Called xBD (x-Building-Damage) it covers the impact of disasters around the globe, like the 2010 earthquake that hit Haiti

**Review: Benefits of ML over regression for clinical prediction**

Journal of Clinical Epidemiology

Performed a life cycle assessment for training several common large AI models

Process can emit more than 626,000 pounds of CO2 equivalent; nearly 5x lifetime emissions of average American car (and that includes manufacture of the car itself)

**Meta-Analysis: Carbon Footprint**

University of Massachusetts

**Review: Neural Machine Translation (NMT)**

Pactera (a global tech company); Insights (consultant to language services industry)

85 percent of (NMT related) AI Projects Ultimately Fail

**Survey: Deep RL For Cyber Security**

Deakin University, Australia and Harvard

Comprehensive review provides foundations for and facilitates future studies on the potential of emerging DRL to cope with increasingly complex cyber security problems

**Survey: GANs**

Dublin City University, Ireland

**Artificial Neuron Using Superconducting Nanowire**

MIT & Colgate University, NY

Introduce engineering concept for a superconducting neuron made from nanowires

Matches energy efficiency of brain - at least in theory - and is the building block of new generation of superconducting neural networks that will be vastly more efficient than conventional computing machinery

**Neural Implant Sends Camera Feed Into Blind People's Brains**

Subjects are able to see points of light on a computer screen using a device called *Orion*

*Orion* implant transmits video images directly to visual cortex, bypassing the eye and optic nerve

First FDA-approved clinical trial of a visual cortical prosthesis

**AI that Can Visualize Objects Using Touch**

The new AI can predict how it would feel to touch an object, just by looking at it

First method that can convincingly translate between visual and tactile signals

**Asynchronously Coded Electronic Skin (ACES)**

Neuro-inspired artificial peripheral nervous system for scalable electronic skins

First to enable many sensors to feed back to single receiver, acting as a whole system

**Collective cognition in the arms of the octopus**

First comprehensive representation of information flow between octopus's suckers, arms and brain

Octopus Arms Make Decisions Without Input From Their Brains

Of the octopus' 500 million neurons, more than 350 million are in its eight arms

**DL of earthquake aftershock patterns**

April 2018 paper "shows" that a NN can predict locations of aftershocks 50% more accurately than classic Coulomb failure stress change methods

Claimed unprecedented accuracy in predicting earthquake aftershocks by using DL

Scathing rebuttal appears in April 2019: "One neuron is more informative than a deep neural network for aftershock pattern forecasting"

Demonstrates that the proposed DNN does not provide any new insight

April 2018 paper "shows" that a NN can predict locations of aftershocks 50% more accurately than classic Coulomb failure stress change methods

Reveals major flaw in original paper: there was overlap in the data used to both train and test the model

Performance of model looks promising because it's being tested using the same data that it was trained on!

**Speech-Driven Facial Animation**

Imperial College, London, UK; Samsung AI Centre, Cambridge, UK

End-to-end system generates videos of talking head, using only still image of person and audio clip containing speech, without relying on handcrafted intermediate features

Generates videos which have: (a) lip movements that are in sync with audio and (b) natural facial expressions such as blinks and eyebrow movements

**Unsupervised Learning Unveils Latent Knowledge**

Lawrence Berkeley National Laboratory, Berkeley, CA

Algorithm analyzes relationships among words in 3.3. million materials-science abstracts; uncovers structure of periodic table, predicts discoveries of new thermoelectric materials years in advance, and suggests as-yet unknown materials

The work does not just establish relationships between words — the researchers also demonstrated how their approach could be used for prospective materials discovery

Suggests that latent knowledge about future discoveries embedded in past publications

As with other ML methods, it is easy to "over interpret" — the algorithm only groups related concepts close together without actually "understanding" or "predicting" anything

The "search" part is less "understanding," and more a "smart search function"

**Dual Video Discriminator GAN (DVD-GAN)**

DeepMind

Scales to longer and higher resolution videos by leveraging computationally efficient decomposition of its discriminator

Creates videos with object composition, movement, even complicated textures like the side of an ice rink

Struggled to create coherent objects at higher resolutions where movement consisted of a much larger number of pixels

**Microsoft invests \$1 billion in OpenAI to create brain-like machines**

Includes developing new supercomputing hardware to try to achieve AGI

**NIST issues draft guideline for developing AI technical standards**

First major step in writing standards that will guide procurement and implementation of AI and ML technologies within the federal government

**National Security Commission on AI Solicits Help**

Independent federal commission: help U.S. determine actions to take to ensure national security enterprise has tools it needs maintain U.S. global leadership

Chartered to produce two reports to Congress

Chaired by Robert Work and Eric Schmidt

**Bill to ban facial recognition from public housing**

First federal bill that looks at what technology landlords can impose on tenant

**DOD Releases Digital Modernization Strategy**

Four strategic initiatives

- Innovation for advantage; Optimization; Resilient cybersecurity; Cultivation of talent

Priorities

- Cybersecurity; AI; Cloud; Command, Control and Communications (C3)

Digital Modernization Goals

- Innovate for Competitive Advantage; Optimize for Efficiencies and Improved Capability; Evolve Cybersecurity for an Agile and Resilient Defense Posture; Cultivate Talent

**Scathing Review of Scotland Yard's Facial Recognition Software**

University of Essex

UK police's facial recognition system has an 81% error rate

Identifies key flaws in the system

- The vast majority of people it flags for the police are not on a wanted list
- Numerous operational failures

**Self-Learning 64-Chip Neuromorphic System**

Intel

Introduced *Pahaki Beach* 8M-neuron neuromorphic system comprising 64 Loihi research chips

Can process information up to 1,000x faster and 10,000x more efficiently than CPUs for specialized applications like sparse coding, graph search and constraint-satisfaction problems

Neuromorphic chips seek to imitate learning ability/energy efficiency of human brains

**Nanophotonic media for artificial neural inference**

University of Wisconsin-Madison

AI made from a sheet of glass recognizes numbers just by "looking"

Glass AI doesn't need to be powered to operate

**Human Replay Spontaneously Reorganizes Experience**

Research probes characteristics of "Replay" in biological and artificial

Answers—in the affirmative—the questions of whether replay can imagine new sequences from whole cloth, and whether sequences shaped by abstract knowledge

Speculate that during rest, brain may explore novel implications of previously learned knowledge by placing an item into an analogy in which it's never been experienced

**Reproducibility in Machine Learning for Health (ML4H)**

MIT, University of Toronto, and New York

ML4H consistently lags other subfields of ML on all measures of reproducibility save inclusion of proper statistical variance

Propose a Reproducibility Taxonomy

- Technical Reproducibility
- Statistical Reproducibility
- Conceptual Reproducibility

**Natural Adversarial Examples**

UC Berkeley; University of Washington; University of Chicago

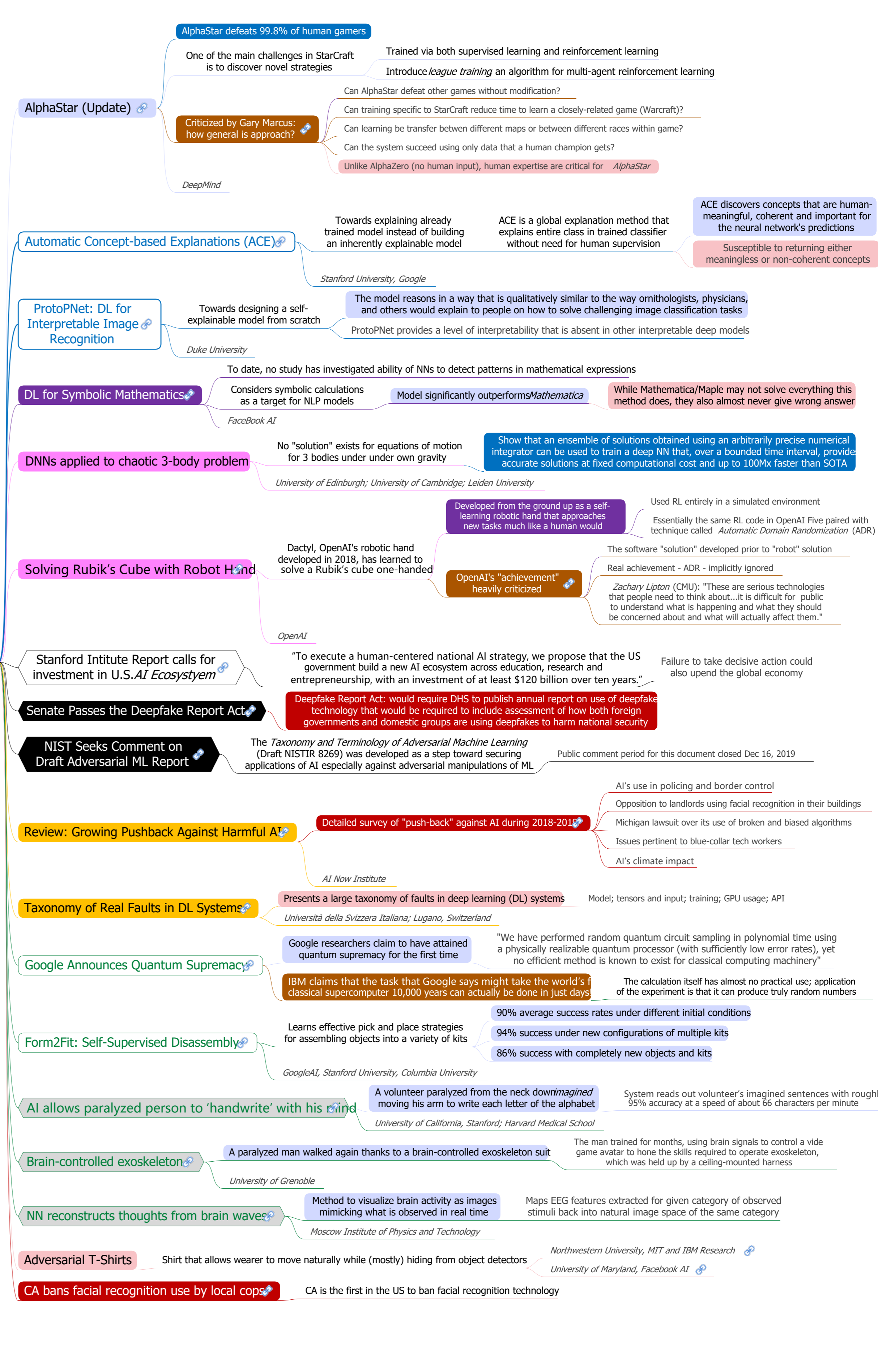
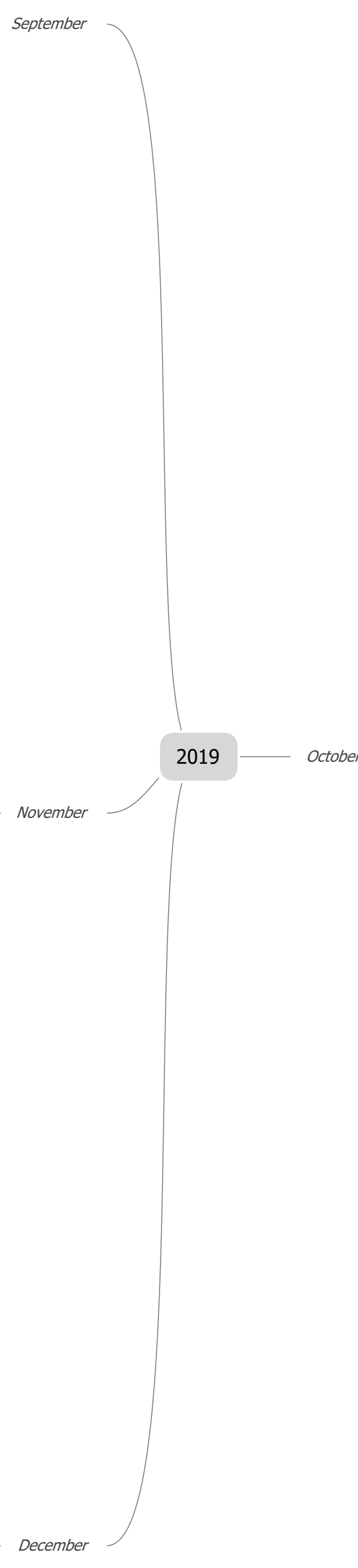
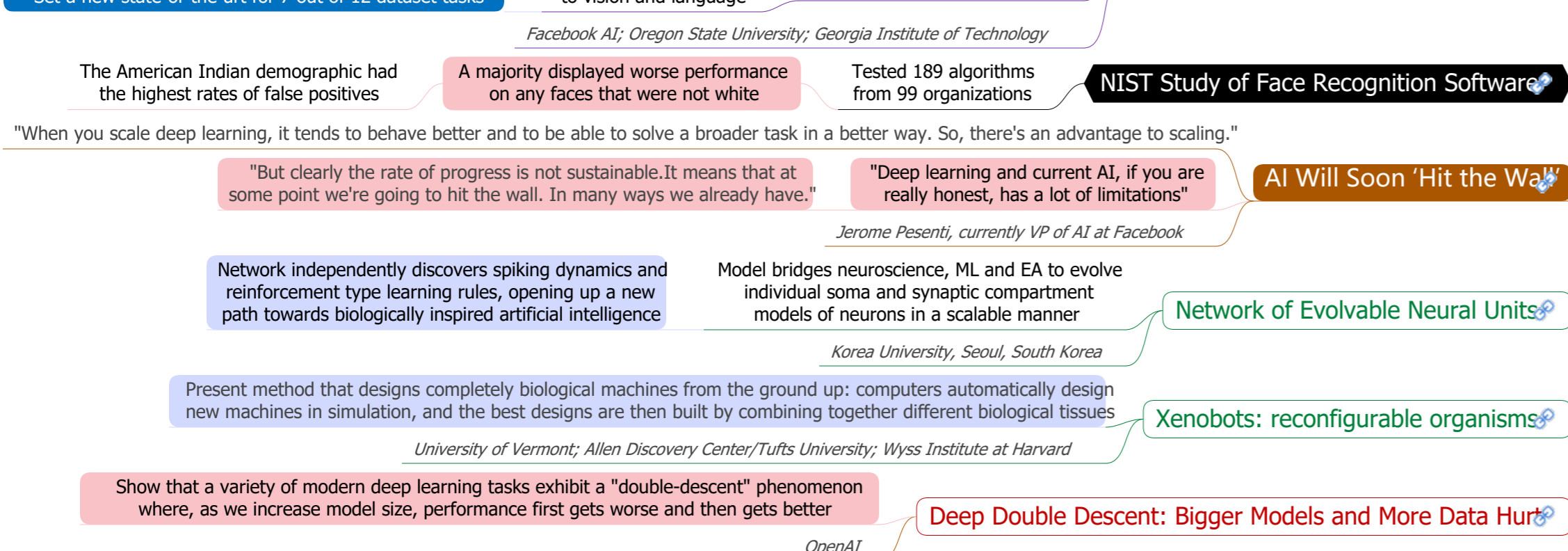
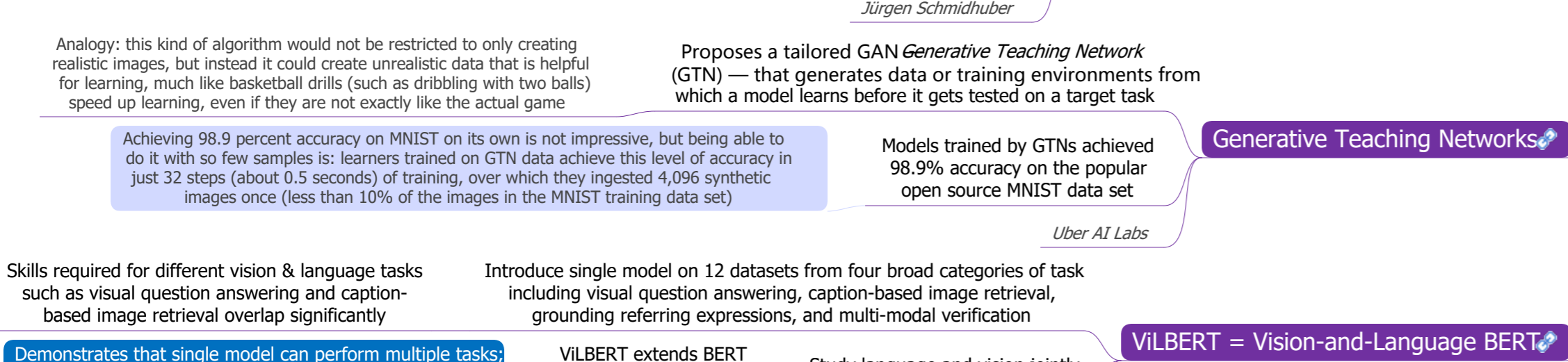
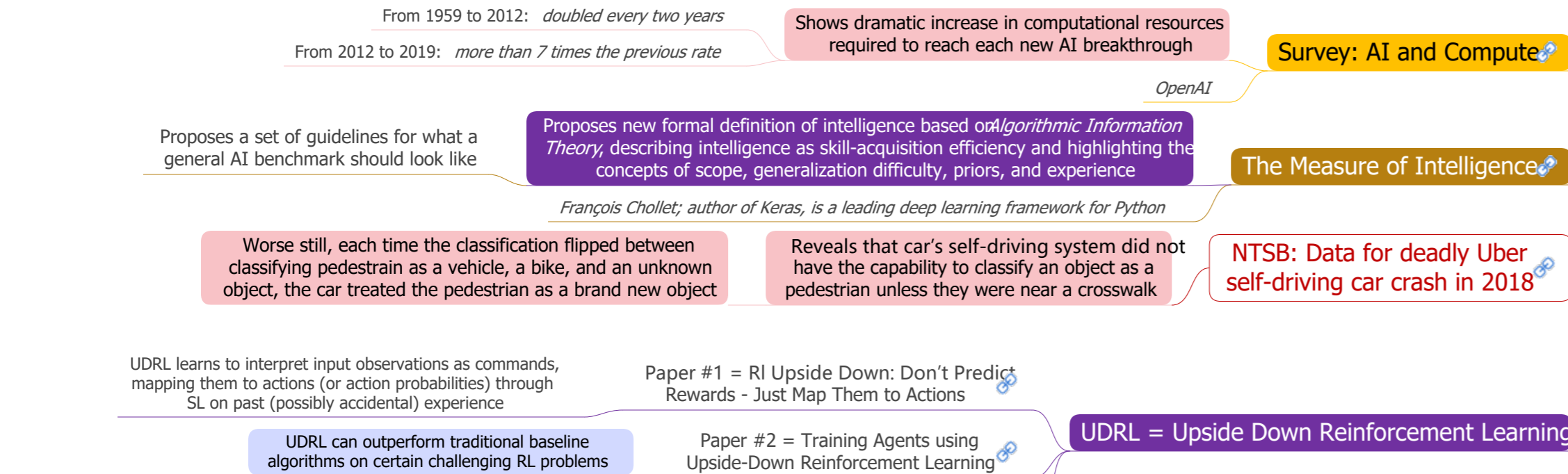
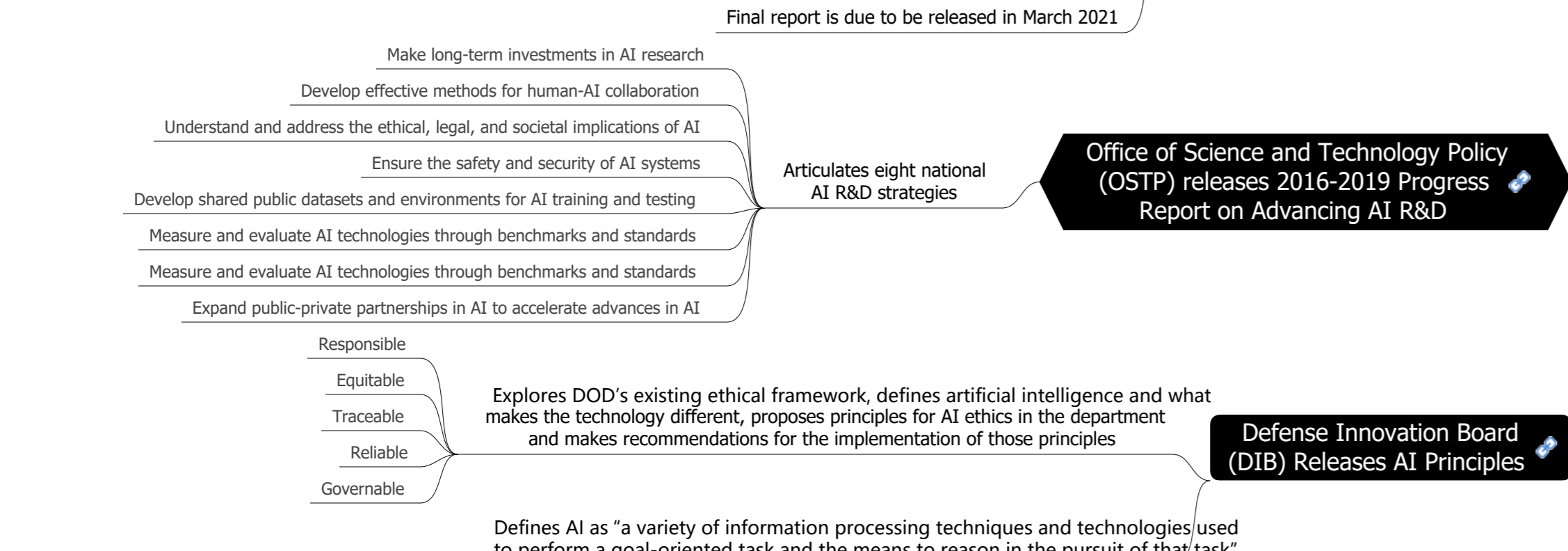
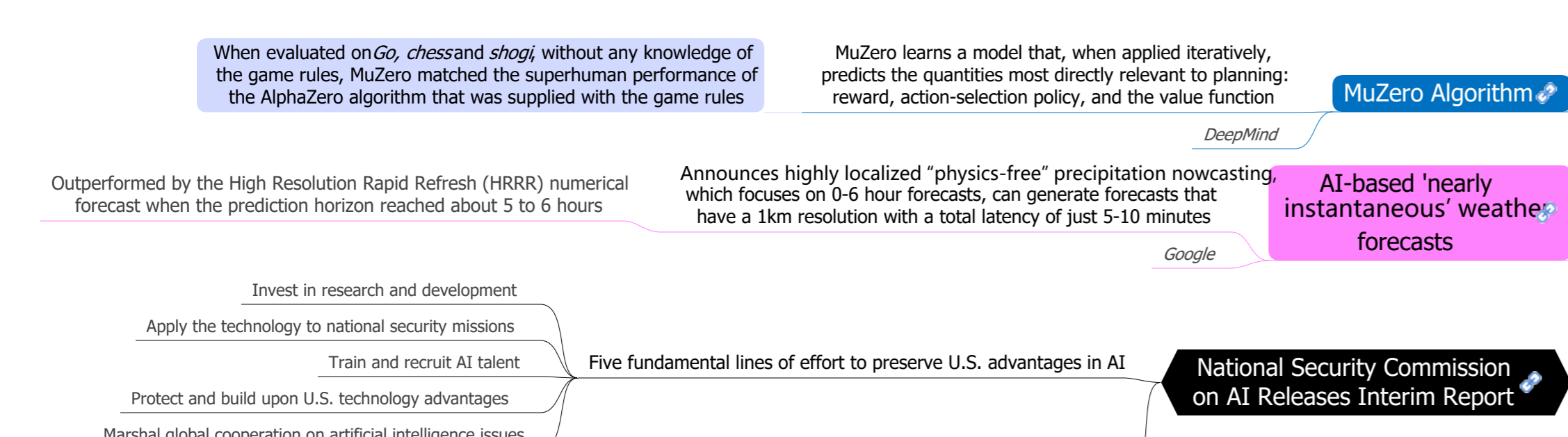
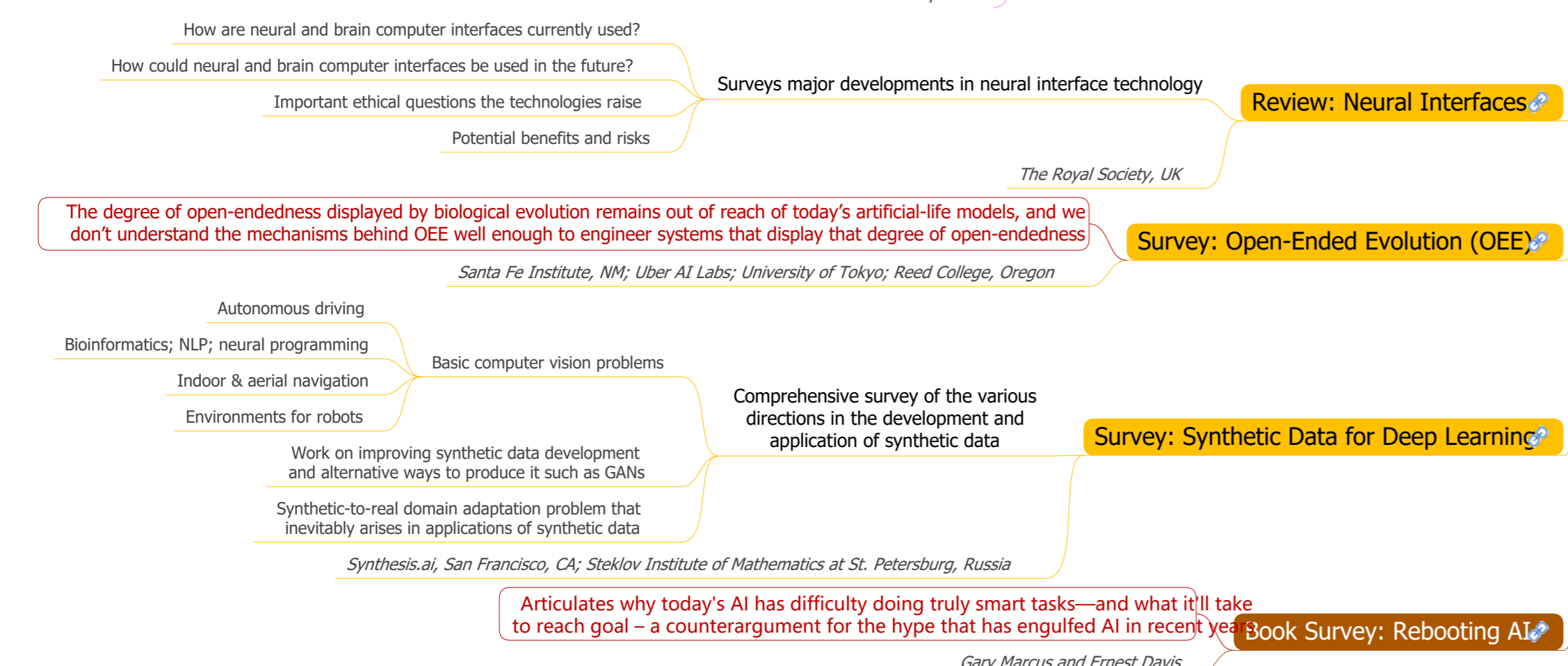
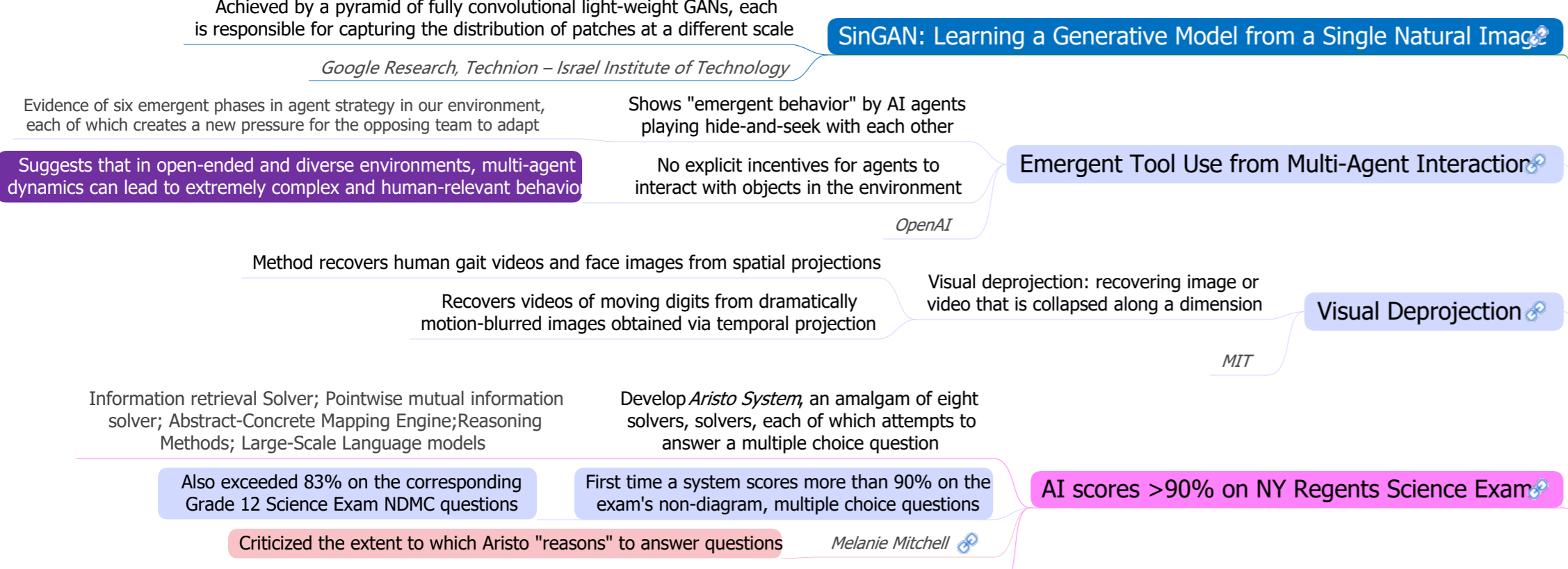
Present real-world, unmodified, naturally occurring examples that cause classifier accuracy to significantly degrade

Shows that vision systems can make unforced errors

AI is essentially missing the wood for the trees

Research suggests that rather than looking at images holistically, considering overall shape and content, algorithms focus in on specific textures and detail

July



### January

- Meena Chat Bot**
  - 2.6B parameter NN trained to minimize perplexity
  - Trained on 341 gigabytes of social-media chatter = 40 billion words (8.5x OpenAI's GPT-2)
  - Meena got an SSA (Sensibleness and Specificity Average) score of 79%, compared with 56% for Mitsuku, a state-of-the-art chatbot that has won the Loebner Prize for the last four years
- Near-perfect navigation without map?**
  - Used large-scale distributed RL algorithm called DD-PPO (Decentralized Distributed Proximal Policy Optimization), which has effectively solved the task of point-goal navigation using only an RGB-D (i.e., depth) camera, GPS, and compass data
  - Facebook AI
- 1st AI-developed drug goes into clinical trials?**
  - Drug development usually takes about 5 yrs to get to trial, but the AI drug took 12 months
  - Drug discovery firm Exscientia; pharmaceutical firm Sumitomo Dainippon Pharma
- Fusion of DL and Combinatorics**
  - Not trying to improve the solvers themselves, but are rather attempting to enable usage of existing solvers in synergy with function approximation
  - One possible approach to introduce combinatorial building blocks into NNs
  - Max-Planck-Institute for Intelligent Systems, Tübingen, Germany; Università degli Studi di Firenze, Italy
- AI Sends 1st warnings of Wuhan virus?**
  - Correctly predicted that the virus would jump from Wuhan to Bangkok, Seoul, Taipei, and Tokyo in the days following its initial appearance
  - Scours foreign-language news reports, animal and plant disease networks, and official proclamations to give clients advance warning of danger zones
  - BlueDot

- Evaluation of AI for breast cancer screening**
  - An example of a potentially great "AI" being used to solve the wrong problem
  - Potential to worsen pre-existing problems such as overtesting, overdiagnosis, and overtreatment
  - How results are used: a deeply flawed medical intervention: cancer screening
  - Google AI
- Insights into how dopamine helps human brain learn**
  - Applied lessons from RL to propose a new theory about the reward mechanisms within our brains
  - Data offers compelling evidence that the brain indeed uses distributional reward predictions to strengthen its learning algorithm
  - DeepMind, Harvard University
- Neuromodulation**
  - Neuromodulation regulates many critical nervous system properties that cannot be achieved solely through synaptic plasticity
  - Results show that - for the three meta-RL benchmark problems considered in this proof-of-concept paper - neuromodulation performs better than classical recurrent NNs
  - University of Liège, Belgium
- San Diego's experiment with FR is a flop?**
  - Since 2012, SD's law enforcement had access to the Tactical Identification System (TACIDS)
  - Not a single arrest or prosecution that stemmed from the program
- Agent57**
  - Agent57 still can't learn to play more than one game at a time; it can learn to play 57 games, but it cannot learn to play 57 games at once
  - Train NN which parameterizes a family of policies ranging from very exploratory to purely exploitative
  - Multiple improvements to deep-Q-learning (DQN); added Never Give Up (NGU) = episodic memory
  - DeepMind
- AutoML-Zero**
  - AutoML-Zero aims to automatically discover computer programs that can solve ML tasks, starting from empty or random programs and using only basic math operations
  - Human-designed components bias the search results in favor of human-designed algorithms
  - Innovation is also limited by having fewer options; you cannot discover what you cannot search for
  - Learning algorithm provided with set of 64 scalar, vector, and matrix operators
  - AutoML-Zero "discovers" linear regression with gradient descent; ReLU = Rectified Linear Unit (a piecewise linear function activation function); 2-layer neural networks with backpropagation; and even algorithms that surpass hand designed baselines of comparable complexity
  - Google Brain / Google Research
- Evolutionary Meta-Learning in Robots**
  - Within a few hours, relying purely on tweaks to current SOTA algorithms, a four-legged robot successfully learns to walk forward and backward, and turn left and right, completely on its own
  - Currently relies on a motion capture system above the robot to determine its location; that won't be possible in the real world
  - Google, Columbia University
- CLEVRER: Collision Events for Video Representation and Reasoning**
  - Shows that "interesting algorithms" can be found using evolutionary search
  - CLEVRER is designed to help evaluate how well AI systems can reason
  - Explores temporal and causal structures behind videos of objects with simple visual appearance
  - Reveals just how bad AI is at reasoning—suggests new hybrid approach requires
  - Outperformed existing models across all categories of questions
  - Introduce Neuro-Symbolic Dynamic Reasoning (NS-DR) model
  - Harvard, MIT CSAIL, IBM-Watson AI Lab, DeepMind
- Enhanced POET: Paired Open-Ended Trailblazes**
  - Enhanced POET is a deliberately open-ended algorithm
  - Poet generates and solves its own challenges, and allows solutions to goal-switch between challenges to avoid local optima
  - Enhanced POET produces a diverse range of sophisticated behaviors that solve a wide range of environmental challenges, many of which cannot be solved through other means
  - Uber AI and OpenAI
- Using AI to Design Chips to Accelerate AI**
  - Algorithm applies deep RL to placement optimization problem
  - Optimizes placement of components on a computer chip to make it more efficient and less power-hungry
  - Performed better than those designed by human engineers
  - Algorithm produced tens to hundreds of thousands of new designs, each within a fraction of a second, and evaluated them using the reward function
  - Google Brain
- PULSE: Photo Upsampling via Latent Space Exploration**
  - Converts 16x16-pixel image of a face to 1024 x 1024 pixels in a few seconds
  - Will not turn out-of-focus unrecognizable image into a clear image of real person
  - PULSE searches a space of high-resolution faces, searching for ones that look most similar to the input image when shrunk to the same size
  - Duke University

### March

- European Parliament Releases Ethics of AI Report**
  - Reviews the guidelines and frameworks that countries and regions around the world
  - Panel for the Future of Science and Technology (STOA)
- ML-Practitioner-Centric AI ethics checklist**
  - Checklist based on six stages of the AI design and deployment lifecycle rather than on a standalone set of ethics principles - from envisioning and defining the system to prototyping, building, launching, and evolving it
  - Addresses perceived disconnect between the focus of the AI ethics community today and the needs of ML practitioners
  - Microsoft
- National AI Initiative Act of 2020 Introduced in House**
  - The bipartisan legislation aims to accelerate and coordinate federal investments in AI, by...
  - Establishing AI institutes to facilitate partnerships between academia, public, private sectors
  - Formalizing interagency coordination
  - Creating an advisory committee
  - Best predictions were not very accurate and were only slightly better than those from a simple benchmark model
  - 160 teams built predictive models for six life outcomes using data from the Fragile Families and Child Wellbeing Study
  - Princeton University
- Review: Measuring AI's Predictability of Life's Outcomes**
  - Dataset consisted of nearly 13,000 data points on over 4,000 families
  - Reviewed a decade of studies that compared the results of the interpretations done by DL algorithms with those done by clinicians
  - Risk toward bias was found to be high in 58 out of 81 studies, "and adherence to reporting standards was suboptimal."
  - Results indicate that many, if not most, claims that AI performs better at interpreting medical images are exaggerated and overhyped, and that the standards of reporting were often poor
  - Imperial College London (published in medical journal BMJ)
- Survey: AI vs Human (Health) Diagnosis**
  - From predictions to understanding
  - Complex transformations of input data
  - Overview of many widely used DL models
  - Google; Cornell University; Schmidt Futures
- Survey: DL for scientific discovery**
  - Unique firing patterns of individual neurons were associated with learning each new word pattern
  - Found that firing patterns of cells that occurred when patients learned a word pair were replayed fractions of a second before they successfully remembered the pair
  - Monitored the electrical activity of thousands of individual neurons as patients took memory tests
  - Brains monitored replaying memories in real time
  - NIH
- Human brain-signals-to-text**
  - Subjects asked to read aloud from set of sentences as team measured brain activity
  - Limited to 250 words; humans typically know about 350K words/use 30K
  - AI's best performance was an average transition error rate of 3%
  - The team tried decoding the brain signal data into individual words at a time, rather than whole sentences, but this increased the error rate to 38% even for the best performance
  - University of California, San Francisco
- Sea Trials Begin for Mayflower Autonomous Ship's 'AI Captain'**
  - Prototype of an AI-powered maritime navigation system ahead of a September 6th venture to send a crewless ship across the Atlantic Ocean on the very same route the original Mayflower traversed 400 years ago
  - IBM; Promare (a U.K.-based marine research and exploration charity)

### February

- ML Helps Discover New Antibiotic**
  - Identified powerful new antibiotic compound
  - In laboratory tests, the drug killed many of the world's most problematic disease-causing bacteria, including some strains that are resistant to all known antibiotics
- Radioactive data**
  - Propose new technique, radioactive data that makes imperceptible changes to this dataset such that any model trained on it will bear an identifiable mark
  - Facebook AI
- New RL method models human behavior**
  - Introduce a toy model of economic competition, and show how reinforcement learning may be augmented with a peer-to-peer contract mechanism to discover and enforce alliances
  - RL is able to adapt if an institution supporting cooperative behavior exists
  - DeepMind
- RIDE: Rewarding Impact-Driven Exploration**
  - Argue that SOTA RL (for multi-player zero-sum games) avoids the notion of cooperation with co-players, a hallmark of the major transitions leading from unicellular organisms to human civilization
  - Exploration in sparse reward environments remains a key challenge of model-free RL (i.e., that effectively relies on "trial and error")
  - RIDE is a novel intrinsic reward for exploration in RL that encourages the encourages agents to explore their environments
  - Agents encouraged to take actions that have significant impact on their representation of the environment state
  - RIDE outperforms state-of-the-art exploration methods, particularly in procedurally-generated environments
  - Facebook AI
- European Commission Releases White Paper on AI**
  - Ecosystem of Excellence: policy framework that sets out measures to align efforts at European, national and regional level
  - Ecosystem of Trust: the key elements of a future regulatory framework for AI in Europe that will create a unique 'Ecosystem of Trust'
  - Describe key dangers and pitfalls in AI development, like bias in datasets, that commercial players have only begun to grapple with
  - Recommended that DoD rely on tools that are transparent, meaning, unlike some types of so-called "black box" neural networks, a technical expert (with permission) could describe the process by which the software reached a specific decision or action
  - Recommend that such tools should be used only within an "explicit, well-defined domain of use"
  - Five principles: Responsible; equitable; traceable; reliable and governable
- DoD Adopt DIBs Principles for Using AI**
  - Recommendations emphasize human control of AI systems
  - Relate local level MAV capabilities to global operations of swarm
  - "Field of robotics and swarm intelligence are both still relatively young, and there remain advances to be made"
  - Frontiers in Robotics and AI
- Survey: micro air vehicles**
  - DL can do single tasks only (play one game, recognize cat photos, etc.) but is not "robust" in unexpected cases (compared with average humans)
  - Proposes hybrid, knowledge-driven, reasoning-based deep-understanding-hybrid approach, centered around cognitive models
  - Gary Marcus
- Next Decade in AI: Four Steps Towards Robust AI**
  - Shows that artificial and biological neurons can be made to communicate bidirectionally and in real time
  - University of Southampton, UK; University of Padova, Italy; ETH Zurich, Switzerland
- Artificial and Biological Neurons Communicate Over Internet**
  - Show that (standard) 300 speller and visual evoked potential BCI spellers can be severely attacked by adversarial perturbations, which are too tiny to be noticed when added to EEG signals, but can mislead the spellers to spell anything the attacker wants
  - Demonstrate neuromorphic computing with an example of facial recognition using ML
  - Huazhong University of Science and Technology, Wuhan, China
- Adversarial attack on brain-computer interface**
  - Development of a nanoscale device that acts like the brain's visual cortex to directly see things in its path
  - University of Central Florida, Orlando
- Artificial biomimetic sight**
  - Bypasses eye and optical nerves: sends signals straight to the brain's visual cortex
  - Uses ML to match the retina's electrical output to simple visual inputs
  - University of Miguel Hernandez de Elche, Spain
- Blind woman plays game sent straight brain?**
  - Towards better understanding relationship between micro-rules and macro-behavior
  - Potentially far-reaching applications: self-organization: leverage connections between convolutional NNs and CA; e.g., self-assembling agents on graphs
  - Google and Allen Discovery Center at Tufts University
- Differentiable Model of Morphogenesis**
  - Introduces first decentralized algorithm with a collision-free, deadlock-free guarantee and validated it on a swarm of 100 autonomous robots in the lab
  - Algorithm's cost is independent of swarm size with respect to computational complexity, memory complexity, and communication complexity
  - Northwestern University, Illinois
- 1st decentralized algorithm for swarming robots**
  - TextFooler can trick NLP systems into misunderstanding text just by replacing certain words in a sentence with synonyms
  - Fooled target models with accuracy of over 90% to under 20%, by changing only 10 percent of the words in a given text
  - Researchers say that tools like TextFooler can help make NLP systems more robust by revealing their weaknesses
  - MIT's CSAIL = Computer Science and Artificial Intelligence Laboratory
- TextFooler**
  - Ask: "Is it possible to attack an RL agent simply by choosing an adversarial policy acting in multi-agent environment so as to create natural observations that are adversarial?"
  - Demonstrate the existence of adversarial policies in zero-sum games between simulated humanoid robots with proprioceptive observations (i.e., relating to stimuli that are produced and perceived within an organism, especially those connected with the position and movement of the body), against state-of-the-art victims trained via self-play to be robust to opponents
  - Adversarial policies are more worrying than attacks on supervised learning models, because reinforcement learning policies govern an AI's overall behavior
  - UC Berkeley
- Attacking Deep RL**
  - Tesla's Mobileye EyeQ3 camera system fooled by subtly altering a speed limit sign on the side of road in a way that a person driving by would almost never notice
  - Camera read the sign as 85 instead of 35, and in testing, both the 2016 Tesla Model X and that year's Model S sped up 50 miles per hour
  - Tesla acknowledged McAfee's findings and said the issues would not be fixed in that generation of hardware
  - Cybersecurity firm McAfee

### April

- Mechanisms for Supporting Verifiable Claims**
  - Multi-stakeholder report by 58 co-authors at 30 organizations
  - Describes 10 mechanisms to improve verifiability of claims about AI systems
  - Third party auditing
  - Bias and safety bounties
  - Sharing of AI incidents
  - Audit trails
  - Interpretability
  - Red teaming exercises
  - Privacy-preserving ML
  - Secure hardware for ML
  - High-precision compute measurement
  - Compute support for academia
- OpenAI Microscope**
  - OpenAI Microscope is a collection of visualizations of every significant layer and neuron of eight vision "model organisms" which are often studied in interpretability
  - Develop a set of recommendations for model interpretation and benchmarking
  - Highlight recent advances in ML to improve robustness and transferability from lab to real-world applications
  - OpenAI, Google Research, University at Oxford, the Centre for the Future of Intelligence, Stanford University, Alan Turing Institute, University of Toronto, Vector Institute, Canada; University of Tübingen, Germany; Max Planck Research School for Intelligent Systems, Germany
- Shortcut Learning in DNNs**
  - Problem: finding a symbolic expression that matches data from an unknown function
  - Combine NN fitting with suite of physics-inspired techniques
  - Apply to 100 equations from the Feynman Lectures on Physics and it discovers all of them, while the "best" previous publicly available software cracks only
  - Department of Physics and Center for Brains, Minds & Machines, MIT
- AI Feynman**
  - Intrinsically-motivated machine learning algorithm (POP-IMGEs), initially developed for learning of inverse models in robotics, can be used to "discover" emergent patterns in self-organized systems
  - System is more efficient than several baselines and equally efficient as a system pre-trained on a hand-made database of patterns identified by human experts
  - University Bordeaux, France
- Intrinsically Motivated Discovery**
  - Technical and Operational Overview of the Next Generation of Autonomous Systems
  - Analysis of various approaches to human control over swarms
  - UNIDIR = United Nations Institute for Disarmament Research
- Swarm Robotics**
  - Presents VCIO model (values, criteria, indicators, and observables) for operationalization and measurement of otherwise abstract principles and demonstrate the functioning of the model for the values of transparency, justice and accountability
  - Proposes context-independent labelling of AI systems, based on VCIO model: seven point scale for transparency, accountability, privacy, justice, reliability and environmental sustainability
  - Labelling approach is unique in the field of AI ethics at the time of writing
  - AIEI = an interdisciplinary European consortium
- An Interdisciplinary Framework to Operationalize AI Ethics**
  - Recommend immediate doubling non-defense AI R&D funding to \$2 billion for FY 2021
  - Accelerate AI applications in DoD
  - Strengthen AI workforce
  - Promote U.S. leadership in AI hardware and SG
  - Improve AI cooperation among key allies and partners
  - Advance ethical and responsible AI
- NSCAI: Releases Recommendations to Congress**
  - National Security Commission on AI (NSCAI) uncovered "concerning developments that amplify the importance of getting AI right for Americans"
  - Title: "AI, Human-Machine Interaction and Autonomous Weapons"
  - Offers thoughts on the public policy challenges presented by the prospect of LAWS
- Office of the Under Secretary of State for Arms Control and International Security Issues Point Paper on AI**
  - U.S. Patent and Trademark Office (USPTO) decided that only "natural persons" - not AI - can be named as inventors, refusing two patents for an AI that created an emergency flashing light and shape-shifting food container
  - Establishing a DoD-wide Responsible AI Subcommittee
- USPTO Denies Patents on Behalf of AI System**
  - Represents "shift" from principles to practice
  - Specific actions include...
    - Using JAIC's Acquisition & Sustainment process to engage how our partners design, develop, and deploy technologies for the DoD
    - Piloting a "Responsible AI Champions" cohort within the JAIC
- JAIC: Describes Plans to Implement AI Ethical Principles**
  - Designed to serve as a primer for DoD officials
  - "Contrary to popular belief, you do not need to understand advanced mathematics or know computer programming languages to be able to answer the above questions accurately and to develop a practical understanding of AI relevant to your organization's needs."
  - "This guide will cover everything that the vast majority of DoD leaders need to know."
- JAIC Releases an 'AI Primer'**
  - Seeks input on cutting-edge testing and evaluation capabilities to support the "full spectrum" of the DoD's emerging AI technologies, including machine learning, deep learning and neural networks
  - Support is needed in the following five areas:
    - Data Set Development/Curation
    - Test Harness Development
    - Model Output Analysis
    - Test Planning, Documentation, and Reporting
    - Testing Services
- JAIC: Issues RFI for New T&E Technologies**
  - Goal is to establish enterprise-wide data governance framework
  - Specifically designed to help JAIC users manage various forms of data at every stage of project development
  - DGC's overall mission is to make sure that the JAIC is properly securing, organizing, storing, and managing its growing collection of data assets on the JCF
  - Tested eight out-of-the-box automatic classifiers, and compared their emotion recognition performance to that of human observers
- JAIC Establishes Data Governance Council (DGC)**
  - Results revealed a recognition advantage for human observers over automatic classification
  - Indicate potential shortcomings of existing out-of-the-box classifiers for measuring emotions, and highlight the need for more spontaneous facial databases that can act as a benchmark in the training and testing of automatic emotion recognition systems
  - University College London; University of Bremen, Germany; Queen's University Belfast, Northern Ireland, UK
- Review: Facial Affect Recognition**
  - 1st look at impact of DL tool in real clinical settings
  - Reveals that even most accurate AIs can make things worse if not tailored to clinical environments in which they will work
  - Google Health
- Evaluation of AI for detecting diabetic eye disease**
  - Goal: to better understand the mechanisms underlying the decision-making processes of NNs
  - Motivation drawn from techniques commonly used in neuroscience studies
  - Identified neuron populations that had different functions
  - Findings imply that specific knowledge stored in a network could be isolated and extracted into a new network, without having to train the new network on the same knowledge
  - Institute of Technologies and Management of Digital Transformation, Institute of Information Management in Mechanical Engineering in Germany
- Investigating role of different neurons in NNs**
  - Two hypotheses - both proven correct
  - That common assumption that learning in the brain is extremely slow might be wrong
  - That dynamics of the brain might include accelerated learning mechanisms
  - Results suggest that adaptation in the brain is significantly accelerated with training frequency
  - This newly observed biological mechanism was used to generalize the conventional "backpropagation" algorithm to include an "acceleration" term
  - Bar-Ilan University, Israel
- Brain experiments reveal mechanisms that outperform common AI learning algorithms**
  - Results suggest potentially deep connections between recurrent predictive neural network models and computations in the brain
  - Model captures wide variety of seemingly disparate phenomena observed in visual cortex
  - From single-unit response dynamics to complex perceptual motion illusions
  - Harvard University; Boston Children's Hospital
- NN trained for prediction mimics biological neurons and perception**

2020

May

June

July

**175 Billion Parameter GPT-3**  
 Shows that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches  
 Noticed pattern wherein gap between zero-shot, one-shot, and few-shot performance grows with model capacity, which suggests larger models are more proficient meta-learners  
 About \$12M (based on public cloud GPU/TPU cost models) = 200x the price of GPT-2  
 Identify some datasets where GPT-3's few-shot learning still struggles  
 OpenAI

**GameGAN**  
 GameGAN: generative model that learns to visually imitate a desired game by ingesting screenplay and keyboard actions during training  
 Goal is to learn a simulator by simply watching an agent interact with an environment  
 First research to emulate a game engine using GAN-based neural networks  
 A realistic copy of the game was created even though there is no code mapping the game's fundamental rules  
 Nvidia, University of Toronto; Vector Institute; MIT

**Plan2Explore**  
 RL agents get specific to tasks they are trained on  
 Agents leverage planning to seek out unexpected future novelty  
 First model that uses planning in learned imaginary latent world model to seek out states in which it is uncertain about what will happen  
 Plan2Explore achieves SOTA zero-shot task performance on a range of tasks, and even demonstrates competitive performance to Dreamer, a SOTA supervised RL agent  
 University of Pennsylvania; UC Berkeley; Google Brain; University of Toronto; Carnegie Mellon University; Facebook AI

**Discovering Physical Laws from Raw Distorted Video**  
 Method for unsupervised learning of equations of motion for objects in raw and optionally distorted unlabeled video  
 MIT

**Decision Points in AI Governance**  
 Overview of 35 efforts to implement AI principles, ranging from tools and frameworks to standards and initiatives that can be applied at different stages of AI development pipeline  
 Start with shared language now exists in the Organization for Economic Co-operation and Development (OECD) AI Principles, which are being leveraged to support partnerships, multilateral agreements, and the global deployment of AI systems  
 Toward operationalizing AI principles  
 The Center for Long-Term Cybersecurity (CLTC) - University of California, Berkeley

**Protection of civilians in armed conflict**  
 The first annual protection of civilians report since 2013 to highlight concerns over killer robots  
 United Nations Secretary-General

**Defense Innovation Unit (DIU) Seeks AI to Detect Shifts in Behavior**  
 Vigilant Keeper: seeks commercial AI solutions to aggregate, structure, and analyze data sets, capture insights, and enable an organization to better support its members who exhibit behavioral changes that may indicate increased vulnerability  
 To prevent mental health issues from becoming a problem on the battlefield, seeks to develop predictive AI that can flag behavior changes in a service member  
 DIU is a team within DoD focused exclusively on fielding and scaling commercial technology across the U.S. military at commercial speeds

**JAIC Unveils "Business Process Transformation" initiative**  
 Business Process Transformation Mission Delivery  
 Part of the JAIC's "year of AI delivery"; i.e., toward fielding new tech not just making policy about it  
 AI-enabled robotic process automation (RPA)  
 Policy Analysis Tool ("Game Changer")

**Review: AI and Efficiency**  
 Since 2012 the amount of computation needed to train a NN to same performance *ImageNet* classification has been decreasing 2x every 16 months  
 Results suggest that for AI tasks with high levels of recent investment, algorithmic progress has yielded more gains than classical hardware efficiency  
 OpenAI

**ML on Graphs: A Model and Comprehensive Taxonomy**  
 Propose a *Graph Encoder Decoder Model (GRAPHEDM)* which generalizes popular algorithms for semi-supervised learning on graphs and unsupervised learning of graph representations into a single consistent approach  
 Fit over 30 existing methods into this framework. We believe that this unifying view both provides a solid foundation for understanding the intuition behind these methods, and enables future research in the area  
 Stanford University; University of Southern California

**EC-Eye: Artificial Eye that "Sees" Like a Human**  
 EC-Eye resembles the size and shape of a biological eye, but with vastly greater potential  
 Up to 10-times more nanowires than biological photoreceptors could potentially be used, allowing EC-Eye to distinguish between visual light and infrared  
 EC-Eye could in theory be connected to an optic nerve to relay information to a human brain, while also improving camera-based eyes currently used on robots  
 Hong Kong University of Science and Technology

**NN trained for prediction biological neurons and perception**  
 Study the emergent properties of a recurrent generative network that is trained to predict future video frames in a self-supervised manner  
 Model is able to capture a wide variety of seemingly disparate phenomena observed in visual cortex  
 That a simple objective—prediction—can produce such a wide variety of observed neural phenomena as demonstrated here underscores the idea that prediction may be a central organizing principle in the brain  
 Harvard University, MA; MIT-IBM Watson AI Lab; Center for Brains, Minds, and Machines (CBMM), Cambridge, MA

**Possible Reproducibility Crisis in Neuroimaging Analysis**  
 Markedly different conclusions about brain scans reached by 70 independent teams highlight the challenges to data analysis  
 Nature (Stanford University)

**Dynamic Stimulation of Visual Cortex**  
 Developed method for drawing symbols – including letter and shapes – directly on the human brain using electrical stimulation  
 Under different paradigms, participants regularly perceived the proper shape in their minds with between 80 and 93 % accuracy  
 For the time being, this technology is "stuck" in the experimental stage  
 Only tested simple shapes such as the letter C and Z  
 UCLA; Baylor University

**World's First Camera Sensors with Built-in AI**  
 Signals acquired by the pixel chip are run through an ISP (Image Signal Processor) and AI processing is done in the process stage on the logic chip, and the extracted information is output as metadata, reducing the amount of data handled  
 Users able to write own AI models to sensors' embedded memory  
 Captures 12-megapixel image and record 4K video at up to 60 frames per second  
 Sony

**Core progress in AI has stalled in some fields**  
 Researchers are more motivated to produce a new algorithm and tweak it until it's state-of-the-art than to tune an existing one  
 There is no clear evidence of performance improvements (of 81 pruning algorithms) over a 10-year period  
 High-water mark for performance of information retrieval algorithms set in 2009  
 6/7 NN-based recommendation systems outperformed by much simpler, non-neural algorithms developed years before  
 Accuracy of ML algorithms has not improved since 2006  
 All ML methods (to solve constrained optimization problems) were found to perform about the same when a simple trick was used to enhance them  
 Science, Vol 368, Issue 6494, published by AAAS = American Association for the Advancement of Science

**Using DL to discover symbolic models**  
 Develop a general approach to distill symbolic representations of a learned deep model by introducing strong inductive biases  
 Apply method to a non-trivial cosmology example: a detailed dark matter simulation  
 Train a Graph NN in a supervised setting, then apply symbolic regression to components of learned model to extract explicit physical relations  
 Discover new analytic formula which can predict the concentration of dark matter from the mass distribution of nearby cosmic structures  
 Find the correct known equations, including force laws and Hamiltonians, can be extracted from the neural network  
 Princeton University; DeepMind; New York University; Flatiron Institute, NY; Carnegie Mellon University

**Unsupervised learning of 3D Images**  
 Method based on an autoencoder that factors each input image into depth, albedo, viewpoint and illumination  
 Leverage symmetry as a geometric cue to constrain the decomposition  
 Outperforms recent SOTA method that uses keypoint supervision for 3D reconstruction on real faces, "while our method uses no external supervision at all."  
 University of Oxford

**Justice in Policing Act**  
 Limitations on use of body camera  
 Video footage may not be subjected to facial recognition or any other form of automated analysis or analytics of any kind  
 Limitations on use of body-worn cameras in conjunction with facial recognition technology for instances, including  
 The use of facial recognition technology only with judicial authorization  
 The use of facial recognition technology only for imminent threats or serious crimes  
 The use of facial recognition technology with double verification of identified faces

**Facial Recognition and Biometric Technology Moratorium Act**  
 Would make it illegal for any federal agency or official to "acquire, possess, access, or use" biometric surveillance technology

**DoD's Office of Inspector General (IG) audit of AI Data and Technology**  
 Reveals gaps and weaknesses in DoD's enterprise-wide AI governance; i.e. JAIC's responsibility  
 Determined that JAIC had not yet developed a department-wide AI governance framework (having passed March 2020 deadline)

**Review: Text Detection and Recognition in the Wild**  
 Highlights challenges affecting text in the wild images that cause existing methods to underperform due to models being unable to generalize to unseen data and insufficient labeled data  
 Examples include in-plane-rotation, multi-oriented and multi-resolution text, perspective distortion, illumination reflection, partial occlusion, complex fonts, and special characters  
 University of Waterloo, Canada

**Survey: Explainable Artificial Intelligence (XAI)**  
 Provides holistic view of the current XAI landscape in deep learning, this paper provides mathematical summaries of seminal work  
 Describe main principles used in XAI research and present the historical timeline for landmark studies in XAI from 2007 to 2020  
 University of Texas at San Antonio

**Brain computation by assemblies of neurons**  
 Hypothesize that assemblies and their operations are involved in mediating higher cognitive functions in humans  
 Introduce *Assembly Calculus (AS)*, which occupies a level of detail intermediate between the level of spiking neurons and synapses and that of the whole brain  
 Develop a mathematical model of neurons and synapses  
 The resulting computational system is shown to be capable of carrying out arbitrary computations  
 Columbia Univ., Georgia Institute of Tech, and Graz Univ. of Technology, Austria

**Fundamental Re-examination of Veracity of Functional MRI data**  
 Hundreds of studies over last decade are based on "common belief" that it is possible to predict an individual's patterns of thoughts and feelings by scanning their brain in an MRI machine as they perform some mental tasks  
 Analysis finds that those measurements are highly suspect when it comes to drawing conclusions about any individual person's brain  
 Psychological Science (Duke University)

**Deep Drone Acrobatics**  
 Introduce NN-based navigation algorithm that enables unmanned quadcopters to pull off acrobatics using onboard sensors  
 Robotics and Perception Group, University of Zurich and Intelligent Systems Lab, Intel

**Energy-Latency Attacks on Neural Networks**  
 Use GA to craft malicious inputs called "sponge examples" = adversarial examples that cause a DNN to burn more energy, take more time, or both  
 Targeting NLP algorithms: increases energy consumption up to 200x ...and latency by up to 70x  
 University of Cambridge, University of Toronto and Vector Institute

**Deepfake Detection Challenge Winner Announced**  
 2,114 participants submitted around 35,000 models trained on its data set  
 Best model able to detect whether video is deepfake with 65% accuracy (on test of 10,000 previously unseen clips)  
 Facebook

**1st Known Case of Erroneous Arrest in the U.S. Due to FR Technology**  
 Robert Williams arrested in Detroit, Michigan in Jan 2020  
 Took his story public in Op-Ed piece in Washington Post on June 24  
 Detroit Police Chief James Craig: "The Detroit Police Department's FR technology misidentifies people about 96% of the time"

**Companies dropping facial recognition (FR)**  
 IBM is Dropping All Facial Recognition Research and Development  
 Amazon bans police from using its FR technology for one year  
 Microsoft won't sell police its facial-recognition technology

**ACM calls for immediate suspension of FR technologies**  
 ACM = Association for Computing Machinery is the world's largest educational and scientific computing society

**Letter to AMS Notices: Boycott collaboration with police**  
 1,400 researchers sign letter to AMS urging researchers to stop working on predictive-policing algorithms and other models  
 AMS = Notices of the American Mathematical Society  
 The AMS itself says that it "has no official position on mathematicians' involvement in providing expertise to law-enforcement agencies, or to companies that do business with such agencies"

**Springer concedes not publish DL paper**  
 Paper: "A Deep Neural Network Model to Predict Criminality Using Image Processing"  
 "This research indicates just how powerful these tools are by showing they can extract minute features in an image that are highly predictive of criminality."  
 Petition: "Crime-prediction technology reproduces injustices and causes real harm"  
 2000+ professors, researchers, and practitioners sign petition

**MIT removes dataset plagued with problems**  
 MIT also urged researchers and developers to stop using the training library, and to delete any copies

**Analogical Reasoning: Scattering Compositional Learner (SCL)**  
 3-layer architecture: object networks, attribute networks and relationship networks  
 Testbed: Raven's Progressive Matrices (RPM)  
 Model discovers compositional representations of objects' attributes (e.g., shape color, size) and relationships (e.g., progression, union)  
 SCL achieves SOTA performance on two RPM datasets, with 48.7% relative improvement on Balanced-RAVEN and 26.4% over another  
 Vector Institute/University of Toronto

**Visual Causal Discovery Network (V-CDN)**  
 Perception module: extracts keypoints  
 Inference module: generates a graph neural network  
 Dynamics module: to predict the future  
 Testbed: simulated environment with fabrics of various shapes and lengths: shirts, pants, and towels  
 Method discovers dependency structures and model the causal mechanisms end-to-end from images in an unsupervised way  
 V-CDN does not solve the grand challenge of causal modeling  
 MIT CSAIL, Nvidia, University of Toronto

**JAIC Shifts Focus to Joint War-Fighting Operations**  
 JAIC initially focused on non-lethal forms of AI  
 Developing AI tools to be integrated into DoD's All-Domain Command and Control (JADC2) program, a SoS approach to connecting sensors to shooters in near-real time  
 JAIC's acting director Nand Mulchandani

**Survey: Drug Discovery with Explainable AI**  
 Summarizes the most prominent algorithmic concepts of explainable AI, and dares a forecast of the future opportunities, potential applications, and remaining challenges  
 ETH Zurich

**The Computational Limits of Deep Learning**  
 Examines computational demands of DL in five prominent application areas: progress in all five strongly reliant on increases in computing power  
 "Continued progress in these applications requires dramatically more computationally-efficient methods, which will either have to come from changes to deep learning or from moving to other machine learning methods"  
 MIT



## **Appendix B: Mindmap of JAIC Milestones 2017-2010**

---

# JAIC Milestones

## 2017

**DoD launches Project Maven**  
Announced by then-Deputy SecDef Bob Work announced in a memo that he was establishing an Algorithmic Warfare Cross-Functional Team, overseen by the undersecretary of defense for intelligence  
Initial focus is to help U.S. Special Operations Command intelligence analysts identify objects in video from small ScanEagle drones

## 2018

Section 238 of the FY 2019 NDAA relates to the Joint AI Research, Development, and Transition activities within the DoD  
Congress meets to establish FY 2019 NDAA  
Jan

Employees sign open An Open Letter To Larry Page, CEO of Alphabet; Sundar Pichai, CEO of Google; Diane Greene, CEO of Google Cloud; and Fei-Fei Li, Chief Scientist of AI/ML and Vice President, Google Cloud  
**Google Employees Resign in Protest Against DoD/Maven Contract**  
May

The memorandum directs the DoD CIO to provide a list of initial NMIs, proposed resourcing plans for both FY 2018 and 2019, and personnel needs by July 27, 2018  
The Deputy Secretary of Defense issues a memorandum establishing the JAIC  
June

The list of initial NMIs was supposed to be launched by September 25, 2018  
The list of initial National Mission Initiatives (NMIs), proposed resourcing plans for FY 2018 and FY 2019, and personnel needs are due  
July 27

The remaining items were not submitted until September 28, 2018  
The FY 2019 NDAA is signed by the President and becomes law  
Aug 13

Within one year, the SecDef shall designate a senior official of the DoD to coordinate activities relating to the development of AI, and define AI for use within the DoD  
The Deputy Secretary of Defense issues a memorandum directing the DoD CIO to establish a JAIC Implementation Team  
Sep 11

The memorandum requests additional personnel resources to support the JAIC Implementation Team  
Goal: to launch a provisional JAIC by Jan 1, 2019, and for JAIC to become fully operational by Oct 1, 2019  
The list of initial National Mission Initiatives (NMIs) related to humanitarian assistance, disaster relief, and predictive maintenance is published  
Sep 28

The JAIC commissions RAND to conduct an independent study to assess DoD's AI posture  
The Senate confirms Lieutenant General Shanahan as the JAIC Director  
Dec 12

## 2019

- Feb 12 **DoD issues its 2018 DoD AI Strategy directing DoD to accelerate the use of AI**
- June 23 JAIC releases open-source natural disaster satellite imagery data set
- June 27 JAIC partners with Singapore's Defence Sci & Tech Agency (DSTA) to develop AI for disaster response
- Aug 13 The FY 2019 NDAA requires DecDef to designate a senior official to coordinate activities relating to AI by this date; and define AI for use within DoD
- Aug 30 JAIC announces that its largest project for FY20 will include "AI for maneuver and fires"**
- Sep 3 DoD seeks 'ethicist' to oversee military AI
- Oct 2 The Deputy DecDef designates the JAIC Director as the senior official responsible for coordinating AI activities within the DoD
- Oct 15 JAIC unveils a public website**
- Nov 7 The JAIC provides the DoD OIG with a draft memorandum to establish an AI Executive Steering Group to provide guidance and oversight of AI policy
- Dec 17 RAND issues report "The Department of Defense Posture for AI: Assessment and Recommendations," which assesses state of AI within DoD
- Dec 23 JAIC issues a Request-for-Information (RFI) for help in applying machine learning to humanitarian assistance and disaster relief efforts
- Dec 25 JAIC issues a Request-for-Information (RFI) for AI developers and drone swarm builders can come together to support search and rescue missions

## 2020

- Jan 16 JAIC announces funding RAND to "explore civil-military views regarding AI and related technologies"
- Jan 31 **Lt. Gen. Jack Shanahan, Director JAIC, announces retirement in summer 2020**
- April 1 JAIC announces its "Responsible AI Champions" for AI Ethics Principles**
- April 13 JAIC issues a Request-for-Information (RFI) for cutting-edge ideas for testing and evaluation capabilities to support the "full spectrum" of DoD's emerging AI technologies, including machine learning, deep learning and neural networks
- April 15 JAIC releases a 20 page AI Primer for DoD officials
- April 30 Establishes Data Governance Council, an enterprise-wide data governance framework**
- May 14 JAIC unveils its "business process transformation" initiative  
Six lines of effort: business administration; acquisition; human capital management; finance and budget; customer relations; and, training and development
- June 4 Nand Mulchandani names acting Director of JAIC**
- June 21 The House Armed Services Committee's (HASC's) Subcommittee on Intelligence and Emerging Threats and Capabilities' version of the NDAA would delegate authority over the JAIC to the deputy SecDef  
The JAIC currently resides under DoD's chief information officer (CIO)
- June 29 DoD Inspector General releases an *Audit of Governance and Protection of DoD AI Data and Technology*  
Reveals a variety of gaps and weaknesses in AI governance across DoD
- July 8 Mulchandani announces shift from applying non-lethal forms of AI (like predictive maintenance) to enabling joint warfighting operations**

## **Appendix C: Mindmap of DARPA AI-related Programs 2017-2010**

---

DARPA

2017

- Lifelong Learning Machines (L2M)** *March*
  - Develop ML frameworks that can continuously apply the results of past experience and adapt "lessons learned" to new data or situations
  - Focus specifically on how living systems learn and adapt and will consider whether and how those principles and techniques can be applied to ML systems
  - Seeks to develop the foundations for systems that might someday "learn" in much the way biological organisms do
- Disruptioneering: Streamlining the Process of Scientific Discovery** *August*
  - Goal: develop new AI-based / mathematical framework for systematic conceptual design of complex mechanical assemblies (e.g., vehicles/aircraft)
- Explainable AI Program (XAI)** *Sep*
- Deep-brain stimulation (DBS)** *Dec*
  - Goal is to use this technique to treat soldiers and veterans who have depression and post-traumatic stress disorder

2018

- Machine Common Sense (MCS)** *March*
  - Two broad strategies
  - The first strategy aims to create a service that learns from experience, like a child, to construct computational models that mimic the core domains of child cognition for objects (intuitive physics), agents (intentional actors), and places (spatial navigation)
  - The second strategy seeks to develop a service that learns from reading the Web, like a research librarian, to construct a commonsense knowledge repository capable of answering natural language and image-based questions about commonsense phenomena
- COMPASS = Collection and Monitoring via Planning for Active Situational Scenarios** *March*
  - Target better understanding of "gray zone"
- Next-Generation Nonsurgical Neurotechnology (N3)** *March*
  - Seeks high-resolution neural interfaces
- Systematizing Confidence in Open Research and Evidence (SCORE)** *June*
  - Aims to develop and deploy automated tools to assign "confidence scores" to different spocial and behavioral science research results and claims
- Artificial Intelligence Exploration (AIE)** *June*
  - AIE Opportunities focus on "Third Wave" theory & applications of AI (i.e., contextual)
- Mosaic Warfare** *July*
  - Shifts warfighting concepts away from a primary emphasis on highly capable manned systems — with their high costs and lengthy development timelines — to a mix of manned and less-expensive unmanned systems that can be rapidly developed, fielded, and upgraded with the latest technology to address changing threats
  - Vision: to transition from kill chains to systems-of-systems adaptive kill webs and tiles
- SHRIMP = Short-Range Independent Microrobotic Platforms** *July*
  - Program seeks innovative designs for robots that measure just a fraction of an inch
  - Goal: develop bots for use in natural and critical disaster scenarios
- Automating Scientific Knowledge Extraction (ASKE)** *August*
  - Goal: an AI that can automatically generate, test and refine its own scientific hypotheses
  - 1st research opportunity under DARPA's *Artificial Intelligence Exploration* program
- Next-Generation Nonsurgical Neurotechnology (N3)** *September*
  - Demo: military pilots control three jets at once via a neural implant
- Subterranean (SubT) Challenge** *September*
  - Designed to explore new approaches to rapidly map, navigate, search, and exploit complex underground environments, including human-made tunnel systems, urban underground, and natural cave networks
- KAIROS = Knowledge-directed Artificial Intelligence Reasoning Over Schemas** *December*
  - Explores how to understand complex events by developing a semi-automated system that identifies, links & temporally sequences their subsidiary elements, identifying participants of complex events & subsidiary elements, & identifying complex event type

2020

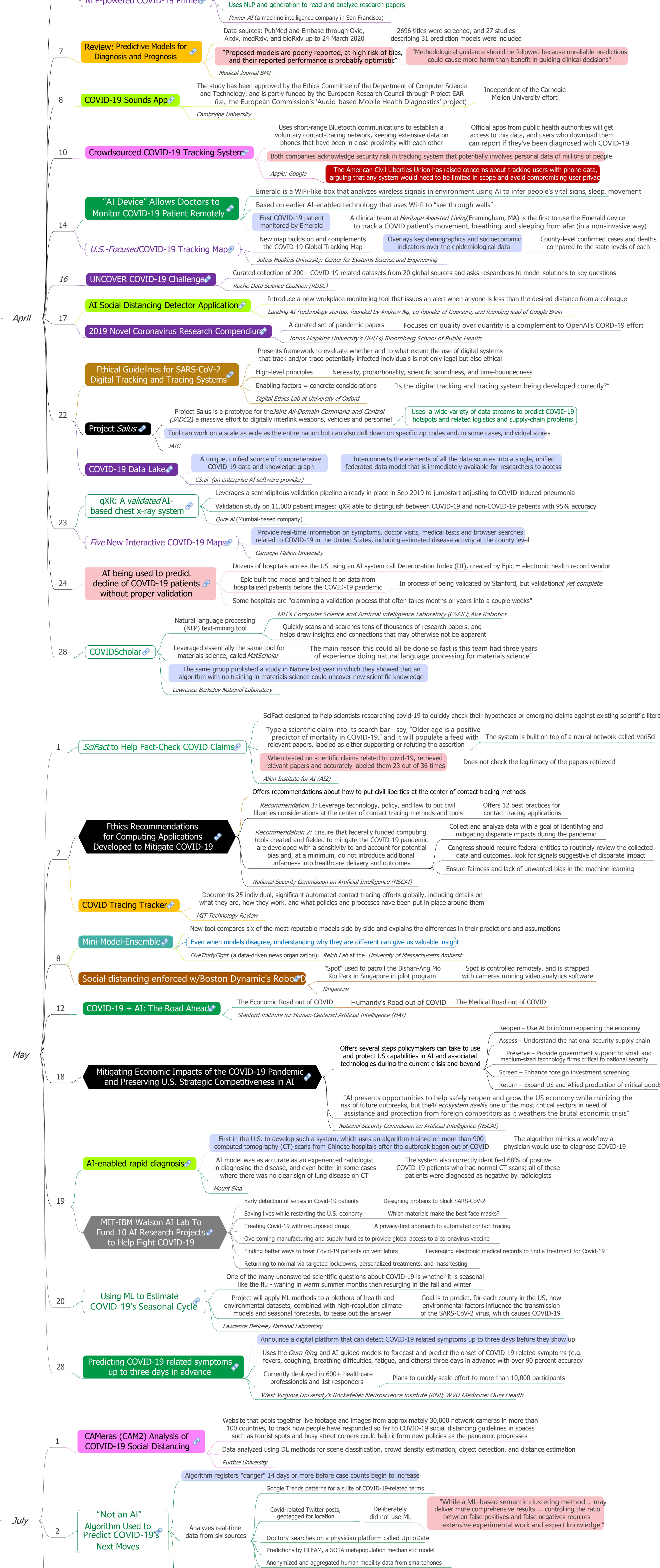
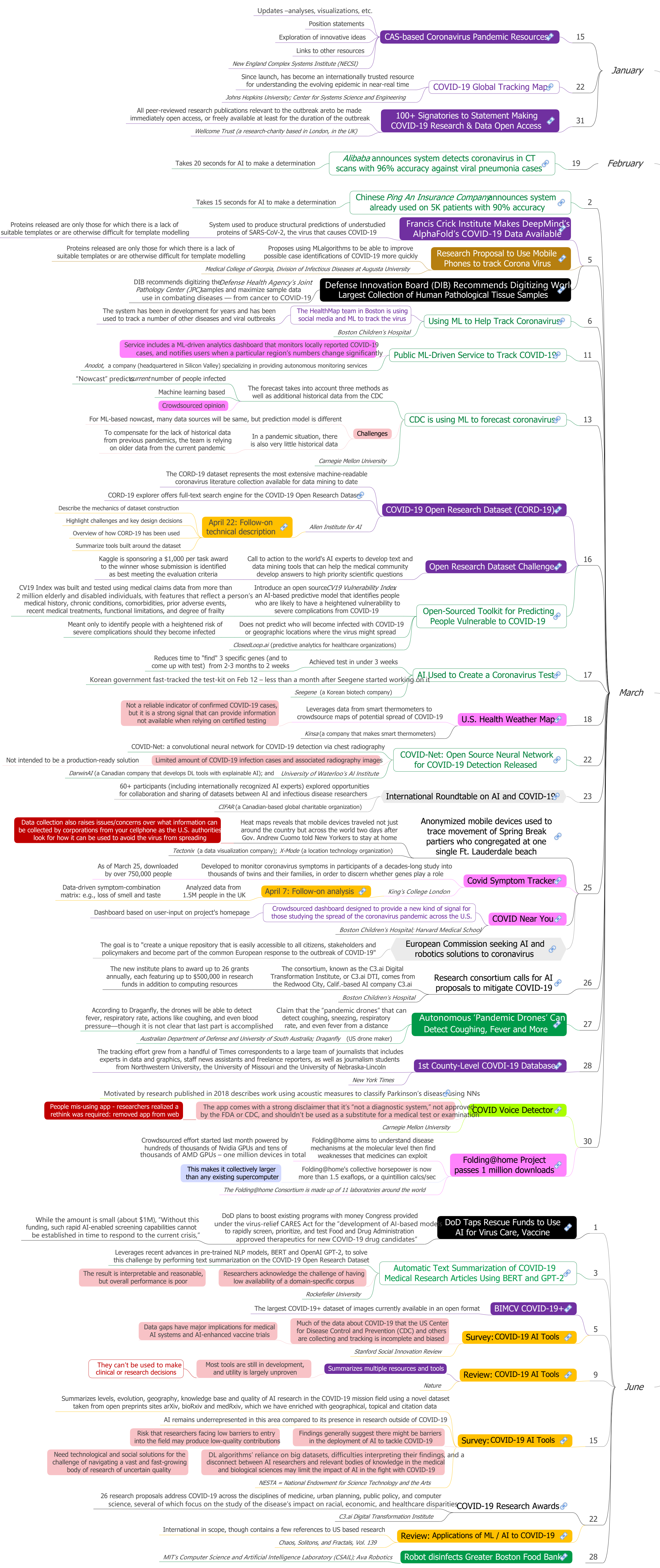
- Habitus** *January*
  - Goal: develop methodologies for creating predictive causal models of local systems based on the input from locals
  - Seeks to capture and make local knowledge available to military operators, providing them with an insider's view to support sound decision-making
- NOMARS = No Manning Required, Ship** *January*
  - Track A (Integrated Seafame Design and Maintenance) will create a framework to evaluate potential design trades against performance requirements
  - The idea is to design a ship from the keel up that will never have a human on board
  - Track B (Enabling Sub-system Technologies) will allow for agile development of relevant subsystem technologies, with a focus on self-adaptive health management for systems relevant to and of similar complexity as that associated with hull, mechanical, and electrical systems
- Gamebreaker opportunity under AIE program** *February*
  - Aims to automate game balance to explore new capabilities/tactics/rule modifications that are most destabilizing to the game or simulation
- Air Space Total Awareness for Rapid Tactical Execution (ASTARTE)** *April*
  - Aims to provide real-time COP of dynamic airspace in most complex and challenging A2/AD environments
  - Does not seek to develop common framework of software / hardware that Joint and Coalition partners would have to acquire
  - Designed for compatibility with existing and future command and control systems (C2) used by the military Services

2019

- Competency-Aware Machine Learning (CAML)** *January*
  - Aims to develop machine learning systems that continuously assess their own performance in time-critical, dynamic situations and communicate that information to human team-members in an easily understood format
  - July 2020 (First milestone) BAE Systems delivered *Mindful.TM* software
- Measuring Biological Aptitude (MBA)** *January*
  - Aims to identify, understand, and monitor in real time the biology that underlies success in specialized roles
  - The research will work backwards from phenotypes — that is, how an individual's fixed genetic code expresses as externally observable cognitive, behavioral, or physical traits — and attempt to establish the biological mechanisms that translate underlying genetic makeup into phenotypic traits
- Intelligent Neural Interfaces (INI)** *January*
  - Seeks to establish proof of concept prototype for 3rd generation AI methods that could improve and expand the application space of next-generation neurotechnology
  - Decision making for sustainment and maintenance of neural interfaces to promote robustness and reliability
  - Modeling and maximizing the information content of biological neural circuits to increase the bandwidth and computational abilities of the neural interface
- Guaranteeing AI Robustness against Deception (GARD)** *January*
  - Aims to develop theories, algorithms and testbeds to help researchers create robust, deception-resistant ML models that can defend against a wide range of attacks, not just narrow, specialized threats
  - Will use a scenario-based framework to evaluate defenses against attacks delivered via sensors, images, video or audio that threaten the physical and digital worlds or the data used to build the ML models
  - Modeling and maximizing the information content of biological neural circuits to increase the bandwidth and computational abilities of the neural interface
  - April 2020 Intel and the Georgia Institute of Technology have been selected to lead GARD
- Science of AI and Learning for Open-world Novelty (SAIL-ON)** *February*
  - Research and develop the underlying scientific principles and general engineering techniques and algorithms needed to create AI systems that act appropriately and effectively in novel situations that occur in open worlds
- Artificial Social Intelligence for Successful Teams (ASIST)** *March*
  - Seeks to develop foundational AI theory and systems that demonstrate the basic machine social skills necessary to facilitate effective machine-human collaboration
- Real Time Machine Learning (RTML)** *March*
  - Seeks to reduce the design costs associated with developing ASICs tailored for emerging ML applications by developing a means of automatically generating novel chip designs based on ML frameworks
- Teaching AI to Leverage Overlooked Residuals (TAILOR)** *April*
  - Aims to to combat Trojan attacks by inspecting AIs for Trojans
  - Obvious defenses against Trojan attacks include securing the training data (to protect data from manipulation), cleaning the training data (to make sure the training data is accurate), and protecting the integrity of a trained model (prevent further malicious manipulation of a trained clean model).
  - In partnership with the U.S. Army Research Office (ARO)
- Air Combat Evolution (ACE)** *May*
  - Aims to increase warfighter trust in autonomous combat technology by using human-machine collaborative dogfighting as its initial challenge scenario
  - Oct 2019: Eight teams selected to compete in the *AlphaDogfight Trials*, a virtual competition designed to demonstrate advanced AI algorithms that can perform simulated within-visual-range air combat maneuvering, colloquially known as a dogfight
- Spectrum Collaboration Challenge (SC2)** *May*
  - Aims to ensure that the exponentially growing number of military and civilian wireless devices will have full access to the increasingly crowded electromagnetic spectrum
  - SC2 teams will develop these breakthrough capabilities by taking advantage of recent advances in AI and ML, and the expanding capacities of software defined radios
- Virtual Intelligence Processing (VIP)** *June*
  - Aims to train AI to process and evaluate information like humans do, and produce actionable results on far smaller datasets than presently done
  - Explore novel / under-explored mathematical and computational "brain-inspired" massively-scalable approaches that have potential to support solutions to real-world DoD problems
  - Also looking for models that can inform the development of future hardware, rather than programs that can run on existing machines
- Intent-Defined Adaptive Software (IDAS)** *July*
  - Seeks to develop engineering methods that separate the problem description — including the problem the software will address, the intent of the software, and its abstract constraints — from any particular instantiation of the software
  - The goal is to drastically reduce the need for manual software modifications, reducing development and maintenance costs and efforts by at least an order of magnitude
- Media Forensics (MediFor)** *August*
  - The goal is an end-to-end media forensics platform capable of detecting manipulations and detail how manipulations were done
- Context Reasoning for Autonomous Teaming (CREATE)** *August*
  - CREATE is investigating "new approaches for autonomous teaming of physically distributed groups of AI enabled systems (multi-agent systems) when there is limited opportunity for centralized coordination."
  - CREATE will develop theoretical foundations of autonomous AI teaming to enable system of heterogeneous, contextually-aware agents to act in decentralized manner to satisfy multiple, simultaneous and unplanned mission goals
- Offensive Swarm-Enabled Tactics (OFFSET)** *September*
  - Envisions swarms of collaborative autonomous systems that provide insights to dismantled troops in urban environments
  - Examples of potential new technologies for integration into the swarm system testbeds include, but are not limited to, swarm sensors, swarm fielding technologies, swarm communications approaches, modular platforms, and mechanisms for swarm manipulation
  - April 2020 DARPA awards contracts to 9 performers to begin work on 5th swarm sprint
- 60th anniversary symposium** *September*
  - Keynote: *Mosaic Warfare* and *Multi-Domain Battle*
  - "Mosaic Warfare: a system-of-systems approach to warfighting designed around 'tiles' of capabilities (i.e., functions: sensors / shooters), rather than uniquely shaped 'puzzle pieces' (i.e., platforms) to be fitted into specific slot in battle plan in order for it to work"
- Artificial Social Intelligence for Successful Teams (ASSIST)** *December*
  - Seeks to develop foundational AI theory and systems that demonstrate the basic machine social skills needed to infer the goals and situational knowledge of human partners, predict what they will need, and offer context-aware actions in order to perform as adaptable and resilient AI teammates

## **Appendix D: COVID-19 & AI Mindmap**

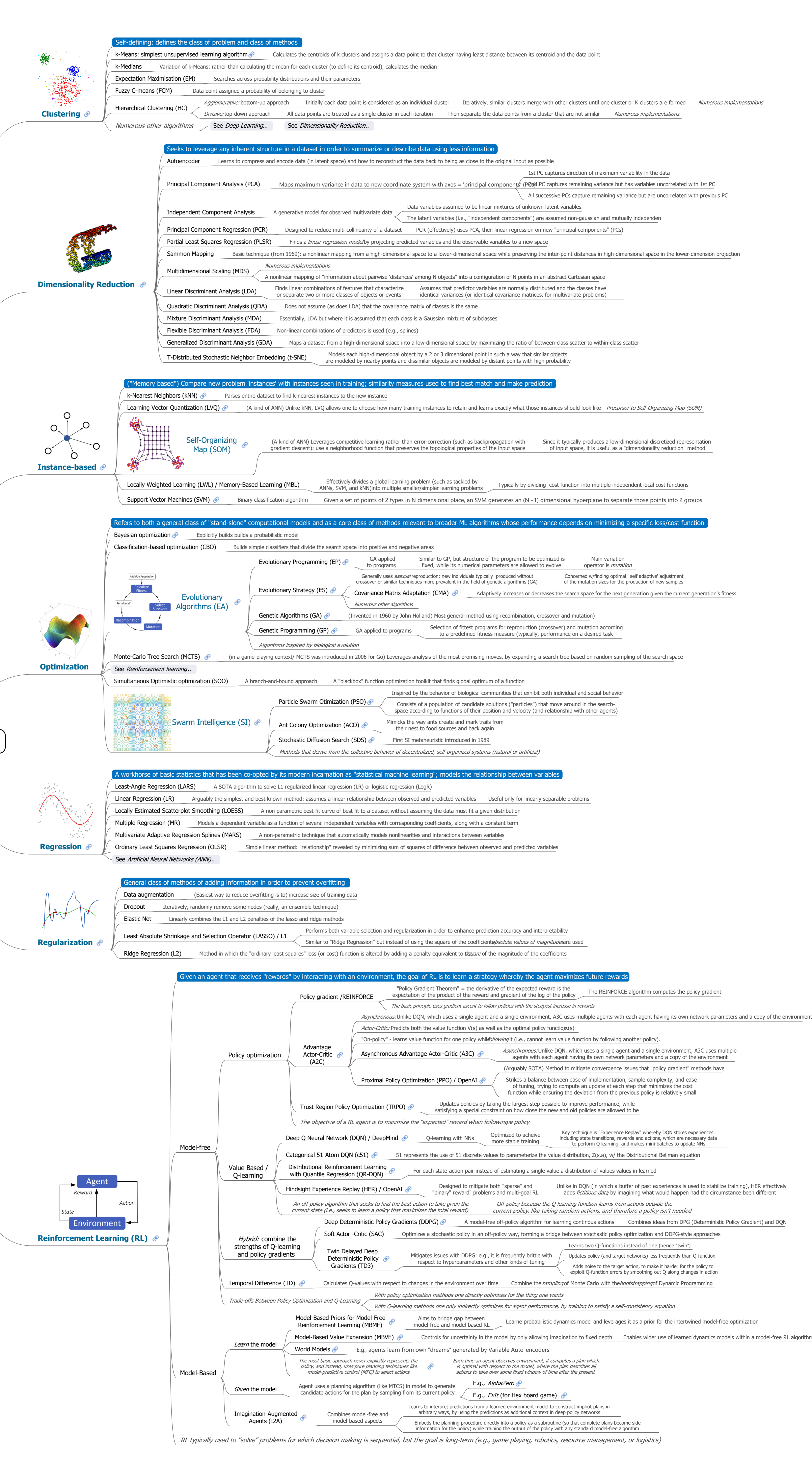
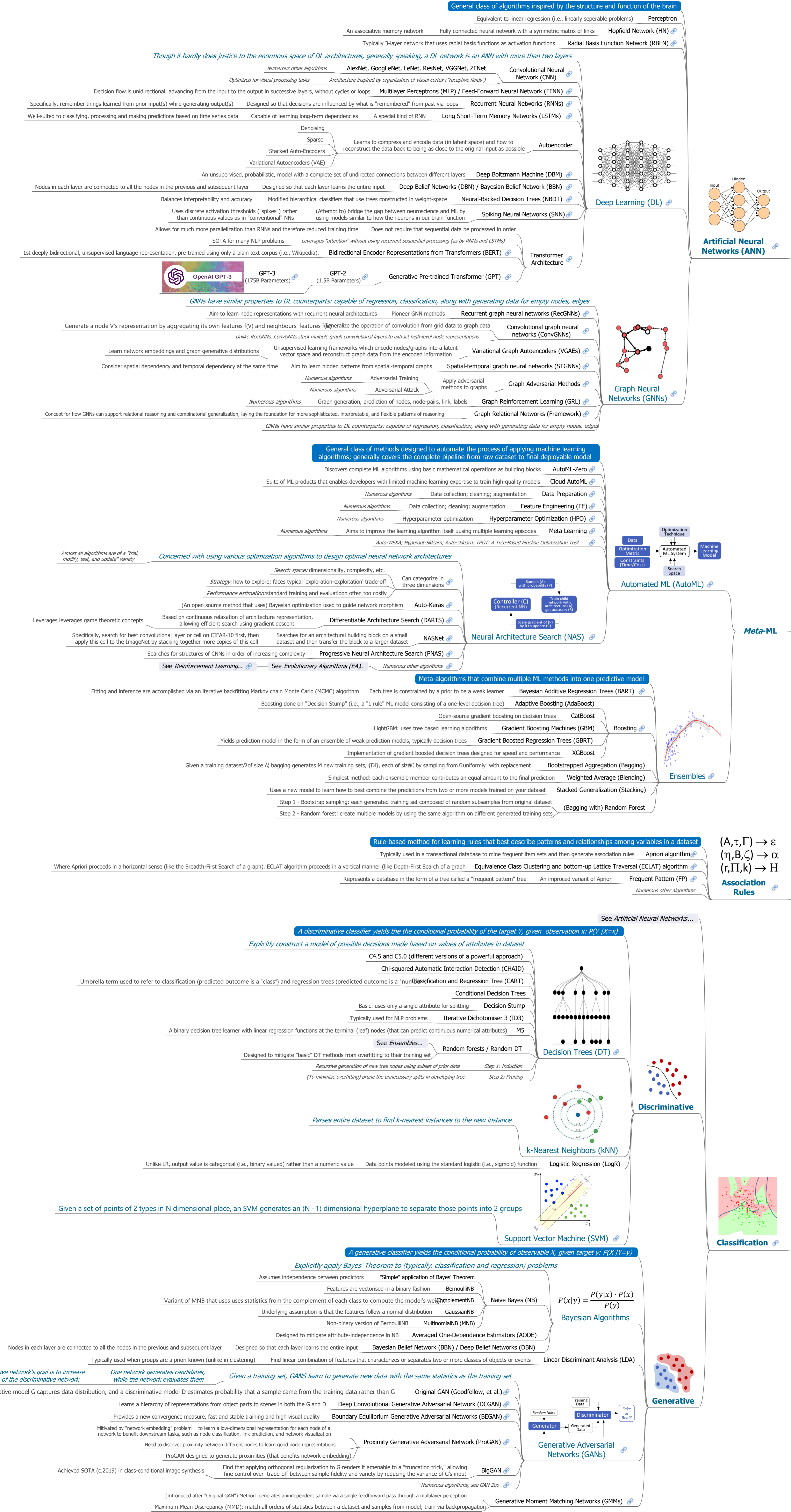
COVID-19 & AI



# **Appendix E: AI/ML Approaches, Methods, and Algorithms Taxonomy**

---

# A Taxonomy of ML Methods and Functions





# **Appendix F: AI Gaps and Limitations**

---

# Fundamental AI Challenges, Limitations, and Vulnerabilities

## Basic Research Challenges

- Fundamental "Devil is in the details" R&D hurdles**
  - Development Timelines
    - Assembling datasets
    - Creating model
    - Training / evaluating / validating model
    - Deployment
  - Both Science and Alchemy
    - Google's Peter Norvig (2016): "ML currently lacks the incrementality, transparency and debuggability of classical programming, trading off a kind of simplicity for deep challenges in achieving robustness."
  - Does not easily distinguish causation from correlation (Lack of) Common Sense
  - (Struggles with) Complex time-varying environments
  - Exploration of vast action spaces (e.g., games) with sparse rewards (RL methods)
  - Human-Machine Teaming (HMT)
    - Communication (Unpredictability of) human component in H-M collaboration
    - Alignment of operator goals/expectations and AI/ML behavior
  - Lifelong Learning
  - Not well integrated with prior knowledge
    - DL methods/"solutions" to a specific "problem" typically self-contained and isolated from other, potentially usefully knowledge
  - Replicability
    - Coding/running new software based on description of a model or method provided in the original publication, and obtaining results similar enough to come to same conclusion
  - Reproducibility
    - Subtle differences in framework implementations, insufficient documentation, hidden hyperparameters, and bugs can cascade and lead to different outcomes
    - Running same software on the same input data and obtaining the same results
  - Limited Transferability (of "solutions" to other problems / domains)
  - (Struggles with) Open ended inference
    - "No DL system currently exists that can draw open-ended inferences based on real world knowledge with anything like human-level accuracy"
  - (Struggles with) Abductive reasoning

## Ethical Challenges

- Law of Armed Conflict
- Accountability and Moral Responsibility
- Human Dignity
- Human Rights and Privacy

## Acquisition Process

- Core Phases
  - Material Solution Analysis (MSA) | Technology Maturation & Risk Reduction (TMRR) | Engineering & Manufacturing Development (EMD) | Production and Deployment (PD) | Operations and Support (O&S)
- Better Buying Power (BBP) Initiative Policy Enhancements
  - Hardware intensive | Software intensive | Incrementally deployed software intensive | Accelerated acquisition | Hardware dominant concurrent with software | Software dominant concurrent with hardware
- Fast-Tracking / Developing Prototypes
  - Defense Innovation Unit (DIU) | Strategic Capabilities Office (SCO)
  - See Test and Evaluation (T&E) and VV&A...

## Integration Challenges

Education | Human-Machine Teaming | Human Resources / Personnel | Culture | Organizational Changes

## Inherent Vulnerability to Adversarial Attacks

- Adversary's Assumed Knowledge
  - Black Box: Adversary has no knowledge about model: uses information about settings and/or past inputs
  - Grey Box: Adversary knows structure and defense strategy, but no knowledge of parameters
  - White Box: Adversary has total knowledge about model
- Goals
  - Confidence Reduction | Targeting Misclassification | Source/Target Misclassification
- Evasion
  - Generative Adversarial Networks (GAN): Adversary adjusts malicious samples so as to force the model to make a false prediction and evade detection
  - Evasion attacks are most common types of attacks
  - Does not assume influence over training data
- Exploratory
  - Model Inversion (e.g., attack infers features used as inputs)
  - Information Inference (e.g., infer whether a given data point belongs to same distribution as the training dataset)
  - Adversary tries to poison training data by injecting carefully designed samples to compromise learning
  - Model Extraction via APIs
  - Attacks do not influence training dataset
- Poisoning
  - Data Injection
  - Data Modification
  - Logic Corruption
  - Unlike in evasion attacks, inputs are modified during training and the model is trained on contaminated inputs to obtain desired output
- All parts of a generic ML data processing pipeline are vulnerable to attack: input data from sensors or data repositories; transferring the data in the digital domain; processing of the transformed data by ML model to produce an output; and action taken based on the output

## Fundamental "Spectre of..." Worries

- AI/ML Development "By the Numbers"
  - Basic computer science, data science, and programming skills being supplanted by "plug in the package" mentality (and use of AutoML routines)
- Ill-equipped "Ecosystem"
  - The overarching architecture, strategy, and vision to evolve and maintain a complex AI-infused ecosystem does not exist
- Decisions are "Too Fast"
  - On Oct 19, 1962, three days into the Cuban Missile Crisis General Curtis LeMay recommended direct military action to President Kennedy. Kennedy rejected the option, and the crisis was resolved diplomatically ten days later
- Incessant Hype
  - Less-than-accurate depictions of research (by those knowledgeable about AI but not technically trained to develop it) exacerbate dichotomy between these groups
- "Manifold Hypothesis"
  - Perturbations to manifold guaranteed to produce erroneous "solutions"
  - The proposition that data naturally forms lower dimensional manifolds in a much higher dimensional embedding space (each manifold represents a different "learned entity")
- Third "AI Winter"?
  - Lack of fundamental ("breakthrough" level) progress in ML research and limitations in computing power may hinder future advancements
- One cannot eliminate spurious samples without sacrificing the model's ability to generate some data we actually want to model

## Autonomous Systems (AS)

- Test and Evaluation (T&E) and Verification, Validation, and Accreditation (VV&A)**
  - Forensic Analysis of Incorrect / Unexpected Behavior(s): The challenge is to find ways to instrument and monitor a system's internal states
  - Fundamental "Micro-Macro" Dichotomy: A (general, omnipresent complex systems theoretic) challenge is to understand the relationship between micro-level specifications and rules and macro-level behaviors
  - Achieving Real-World-Level Fidelity of Training Data: The challenge is to achieve a "realistic enough" mapping between "reality" and a system's internal representation of that reality to support all mission requirements (JHU/APL's Range Adversarial Planning Tool (RAPT))
  - (Adaptive) CONOPS Development: The challenge is to account for a priori unanticipated interactions between autonomous system(s) and humans as mission(s) unfold
  - Trust: Trust not an innate trait of the system; rather, a relative measure of how a human operator(s) experiences and perceives the behavioral pattern of a system
    - Barriers to establishing trust
      - Sensing and thinking disconnect between AS and humans
      - Lack of situational awareness (among disparate environments)
      - Predictability and directability: systems must be able to both communicate relevant information an understandable way and anticipate events as they might unfold
      - Commensurability / alignment of human-machine goals
      - Human-machine interfaces powered by context-aware NLP algorithms
      - Life long ML that allows systems to adapt to conditions and contexts beyond those for which original VV&A was performed (and for which may no longer be valid)
  - Elevated Safety Concerns and Asymmetric Hazard: T&E/VV&A of autonomous systems - and, therefore, the safety of test environments - necessarily combines a mix of operator decisions and (not always predictable) system behavior
  - AS-Specific Vulnerabilities: When operating on their own, autonomous systems are vulnerable to modes of attacks that would be less of a concern for human operators (e.g., cyber, physical, and ML-specific) (See Inherent Vulnerabilities...)
  - Emergent Behavior: DoD Directive 3000.09 specifically warns against the possibility of "unanticipated emergent behavior resulting from the effects of complex operational environments on autonomous or semi-autonomous systems."
  - Post-Trained Learning: Systems may continue to learn after training (either to adapt to changing conditions during actual missions and/or to incorporate additional data post deployment); TE/VV&A must accommodate continually changing behaviors
  - VV&A of Training Data (See Inherent Vulnerabilities...)
- Swarms**
  - Resilience: How well does the system continue to perform assigned mission or tasks in degraded conditions?
  - Predictability vs. Control: Tradeoff between predicting how a system will interpret instructions (implement a given task, generally behave...) and being unable to control it (giving the system the ability to find "novel" solutions")
  - Influence: How much control of - and ability to intervene in - system's behavior does operator/team-mate have?
  - Modeling & Simulation: Tradeoff between "augmenting" existing datasets and introducing unwanted uncertainties and/or bias
  - Speed (of Control): How quickly does the system (e.g., swarm) process and put-into-effect commands given to it?
  - (Swarm-specific) Technological sophistication: What is the tradeoff of sophistication of individual parts and size of swarm?
- Lack of a universally accepted conceptual framework**
  - Multidimensional Categorizations
    - Distinguish among the number of systems required for a given task
    - Segregate functions according to the nature of the tasks involved
    - Emphasize the degree of objectivity required in the decision-making process
    - Evaluate functionality in the context of situational difficulty
    - Focus on the degree of complexity required of human-control
  - Multidimensional Sets of Measures of complexity
    - 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 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
  - Prototypes
    - Autonomy Levels for Unmanned Systems (ALFUS) developed by the National Institute of Standards and Technology (NIST)
    - Autonomous System Reference Framework (ASRF) introduced in DoD's Defense Science Board's (DSB's) 2012 report on autonomy
- Interoperability**
  - General Challenges
    - Multiple networks and gateways
    - Unmet or poor interface standards
    - Platform-centric requirements
    - Poor training and underdeveloped CONOPS
  - Inherent AS Challenges
    - Potentially fewer communication opportunities because of autonomous operation
    - Lack of a human operator to override potential problems
    - Managing risk from a more agile acquisition process
- Intensely Data-Hungry / Data-Driven**
  - Magnifies "garbage in, garbage out" problem
    - Datasets for military organizations, training facilities, platforms, sensor networks, weapons, etc. were initially not designed for ML-purposes
    - Possible mitigations
      - Generative adversarial networks
      - Transfer learning: open research
      - Modeling & simulation
  - Incomplete, not-well-curated, and/or biased datasets: The fusion of multi-domain/multi-source datasets (possibly in real time) requires a new generation of data fusion / algorithms and architectures
  - Fundamental "Fat Tails" Problem: Sparse / not completely representative and/or biased datasets
  - Using M&S to Augment Existing Datasets
- (Un-)Predictability**
  - Brittleness of Image Classification Systems
  - Inherent / irreducible(?) potential for accidents and mistakes
  - Inherent, unavoidable, and unanticipated emergent behavior in complex systems-of-systems (AI-v-AI interactions)
- Explainability, Interpretability, Understandability**
  - Goal (of "explaining")
    - Goal Alignment (between operator and AI)
    - Task Alignment
  - Scope
    - Local: "explain" individual features of single instance of input data x from the data population X
    - Global: provide insight into model as a whole; "understand" attributions for an array of input data
  - Method
    - Data-centric / backpropagation: core logic dependent on gradients backpropagated from output prediction layer back to input layer
    - Parameter-centric / perturbative: logic is dependent on random or selected changes to features in the input data instance
    - Self-explaining: provide "explanation" of each decision with confidence levels for both decision and explanation
  - Usage
    - Intrinsic: explainability is baked into architecture itself and is generally not transferable to other architectures
    - Post-Hoc: algorithm is not dependent on model architecture and can be applied to an already trained NN
  - Fundamental tradeoffs between explainability/interpretability and performance/accuracy

## **Appendix G: Neuro-Lego World**

---

# A Timeline of Deep Learning Architectures Re-Imagined as an evolution of problem-inspired "Building Blocks" in a "Neuro-Lego-World"

**1958** **Perceptron**

**Can we leverage a basic understanding of how the brain works to develop an AI?**

● = Output neurons    ● = Input neurons

Most basic building block of neural networks  
Consists of a simple input and single output cell  
Constrained to "solving" only linearly-separable problems

**Introduce neurons and synapses**

Frank Rosenblatt, "The Perceptron—a perceiving and recognizing automaton"

**1982** **Hopfield Network (HN)**

**Can a network of artificial neurons and synapses act as a "memory"?**

Neurons are binary-valued threshold units; i.e., value determined by whether or not their input exceeds a given threshold

Leverages basic physics concepts (e.g., "relaxation" to minimal energy states): associate an "energy" with each network state

Training consists of lowering energy to its "minimum" value (associated with a desired "memory" state)

Since HNs are guaranteed to converge to a *local* minimum, the "memory" may represent an undesired pattern

**Connect each neuron to all other neurons**

John Hopfield, "Neural networks and physical systems with emergent collective computational abilities"

**1986** **Boltzmann Machine (BM)**

**Can a Hopfield Network do more than just retrieve memories, like solve problems?**

● = Hidden neurons

Labels some neurons as "input neurons" and leaves others as "hidden neurons"

Neurons are the same binary units as in Boltzmann machines and are all completely connected

The goal is not to retrieve memories but to learn representations of input data

Instead of minimizing energy function (as in HNs), training minimizes error between input data and the representation achieved by the hidden neurons (and their weights)

**Separate neurons into input and hidden layers**

Geoffrey Hinton and Terrence Sejnowski, "Learning and relearning in Boltzmann machines"

**1986** **Feed-Forward Neural Network (FFNN)**

**Can Perceptrons be used to classify data that is not linearly separable?**

Contains an input layer, multiple hidden layers, and an output layer

Use backpropagation to iteratively update parameters desirable performance is achieved

Proven to be universal function approximators

**Add hidden layers**

David Rumelhart, Geoffrey Hinton, and Ronald Williams, "Learning representations by back-propagating errors"

**1988** **Autoencoder (AE)**

**Rather than storing and retrieving information, can we compress it?**

● = Encoder / decoder neurons

Symmetrical input and output layers

Layers closest to input *encode* data

Middle (bottleneck) layers are smallest to force compression

Layers closest to output *decode* data

**Add bottleneck between input and output**

Bourlard and Kamp, "Auto-association by multilayer perceptrons and singular value decomposition"

**1990** **Recurrent Neural Network (RNN)**

**What if the form of the data is sequential, as in text-based input?**

● = Recurrent neurons

FFNNs "remember" things but only what they learned during training

Add a loop in the network that passes prior information forward

Loops effectively step the network through sequential input data while maintaining the data in the form of "hidden layers" between steps; i.e., a rudimentary "working memory"

RNNs may be viewed as multiple copies of an FFNN executing in a chain

**Add feedback loops to remember the past**

Jeffrey Elman, "Finding structure in time"

**1997** **Long Short-Term Memory (LSTM)**

**What if an RNN's "working memory" is too short; what if we need more 5 or 10 time-steps between input events and target signals?**

● = Memory neurons

LSTMs are essentially RNNs that are able to remember information an arbitrarily long time

*Input gate*: determines how much information from previous layer gets stored in cell

*Output gate*: determines how much information the next layer gets to know about the state of a given cell

Forget gate: tells the cell state which information to forget

Inspired mostly by electrical circuit design, and not so much biology

Endow each neuron with a memory cell and three gates: *input, output* and *forget*

**Add circuits to regulate information storage**

**1997** **Long Short-Term Memory (LSTM)**

**What if a solution to a problem requires knowing not just past states but also future states?**

Connect two hidden layers of opposite directions to the same output

Output layer receives information from the past and future states simultaneously

**Add feedback loops to inject future states**

Bidirectional Recurrent Neural Network (BiRNN), Bidirectional Long Short-Term Memory (BiLSTM)

Schuster and Paliwal, "Bidirectional recurrent neural networks"

Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory"

**1998** **Convolutional Neural Network (CNN)**

**Can we mitigate the computational cost associated with applying FFNNs to the inherently high dimensionality of images?**

**Use convolutional and pooling layers to reduce dimensionality of image**

● = Kernel neurons    ● = Convolution / pool neurons

A convolution is a *kernel*, i.e., an x-by-x weight matrix such that, as it slides around an image, yields a weighted sum of the pixels underneath the kernel

Convolutional layers typically shrink in size as they become deeper (easily divisible factors of input)

*Pooling filters* out details; e.g., max pooling, whereby, say, a 2 x 2 pixel patch is filtered to pass forward the pixel that has the greatest "redness"

LeCun, et al., "Gradient-based learning applied to document recognition"

**2004** **Echo State Network (ESN)**

**Can we speed up the learning process for RNNs and make it work better for dynamical time-series data?**

**Pool all hidden layers into one "reservoir" and randomize synapses**

RNN architecture but with a very sparsely connected hidden layer (typically, about one-percent connectivity)

Neuronal connectivity and weights are randomly assigned

The input layer is used to prime the network

Output layer acts as an observer of activation patterns that unfold over time

The input signal is connected to a fixed (non-trainable) and random dynamical system (the reservoir = hidden layers), creating a higher dimension embedding

Basic element of *reservoir computing* architectures

Herbert Jaeger and Harald Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication"

**2010** **Deconvolutional Neural Network (DNN) (or Inverse Graphics Network, IGN)**

**Instead of classifying images, can we generate them?**

**Create the opposite of a CNN: apply deconvolution**

Effectively a reversed CNN; e.g., one that takes as input, say, an image of a dog and classifies it

Reverse flow: an input vector representing a given class (e.g., "dog"), is mapped to an image of a dog

Can also be used for direct image deconvolution (i.e., removing noise, artifacts, and other degradations)

Zeiler, et al., "Deconvolutional networks"

**2014** **Generative Adversarial Network (GAN)**

**Instead of enhancing input images (or producing noisy exemplars), can we generate an unlimited set of novel class-specific images?**

**Set up a competition between two networks**

Noise → DNN → Fake Image

Real Image → CNN → Real or Fake?

One network, the "Generator" (a DNN) generates images

The Generator's goal is to minimize the "Discriminator's" performance

A second network, the "Discriminator" (a CNN) decides whether an image is real or fake

The Discriminator's goal is to maximize its ability to identify real/fake images

Goodfellow, et al., "Generative adversarial nets"

**2015** **Residual Network (ResNet)**

**What if there are so many layers that backpropagation does not work (i.e., "vanishing gradients")?**

**Skip some connections**

Design architecture deliberately propagates signals across skipped layers

Learning is enhanced since the fewer number of layers reduces the impact of vanishing gradients

Partly motivated by "skipped layers" in the biology of the human brain

Kaiming He, "Deep Residual Learning for Image Recognition"

**2015** **Attention Network (AN) (includes Transformer Architectures)**

**Can we mitigate CNN's lack of spatial invariance to input data (in a computationally efficient manner)?**

**Add module that can manipulate data within the network**

New module ("Spatial Transformer") is inserted directly into an existing convolutional architecture

An attention mechanism mitigates information decay by separately storing previous network states and switching attention between the states

The hidden states of each iteration in the encoding layers are stored in memory cells

Decoding layers connected to encoding layers, and receive data from memory cells filtered by attention context

A filtering step adds context for decoding layers focusing attention on particular features

Jaderberg, et al., "Spatial Transformer Networks"

**2017** **Capsule Network (CapsNet)**

**Can we leverage the spatial relationships among parts of images (that CNNs effectively ignore; i.e., the "Picasso Problem")?**

**Add capsules**

● = Capsule neurons

Each capsule has a logistic unit to represent the presence of an entity and an N-by-N pose matrix which learns to represent the relationship between an entity and the viewer

Also enhances a network's resistance to adversarial attack

Note: to date, and despite its intuitive appeal, there have been few proven results using this concept

Hinton, et al., "Capsule Networks (CapsNet)"

**2017** **Automated Machine Learning (AutoML)**

**Can we mitigate the increasingly intractable combinatorial search problem of finding an "optimal" NN design for a given problem?**

**Bootstrap design of machine learning models by using machine learning**

Use a controller NN (e.g., RNN or LSTM) to propose a "child" model architecture which is trained and evaluated for the given problem

Use feedback from evaluation to inform controller to improve its proposals for next round

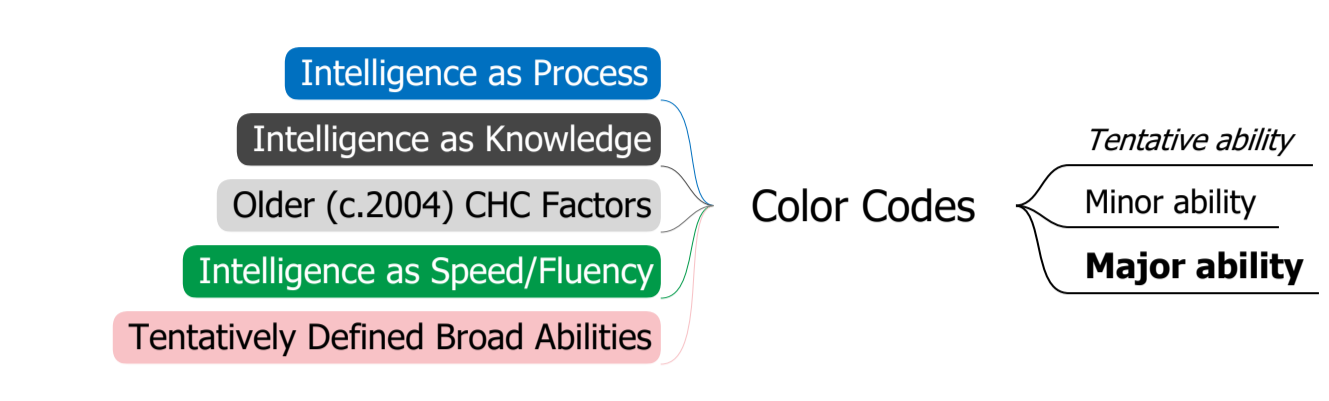
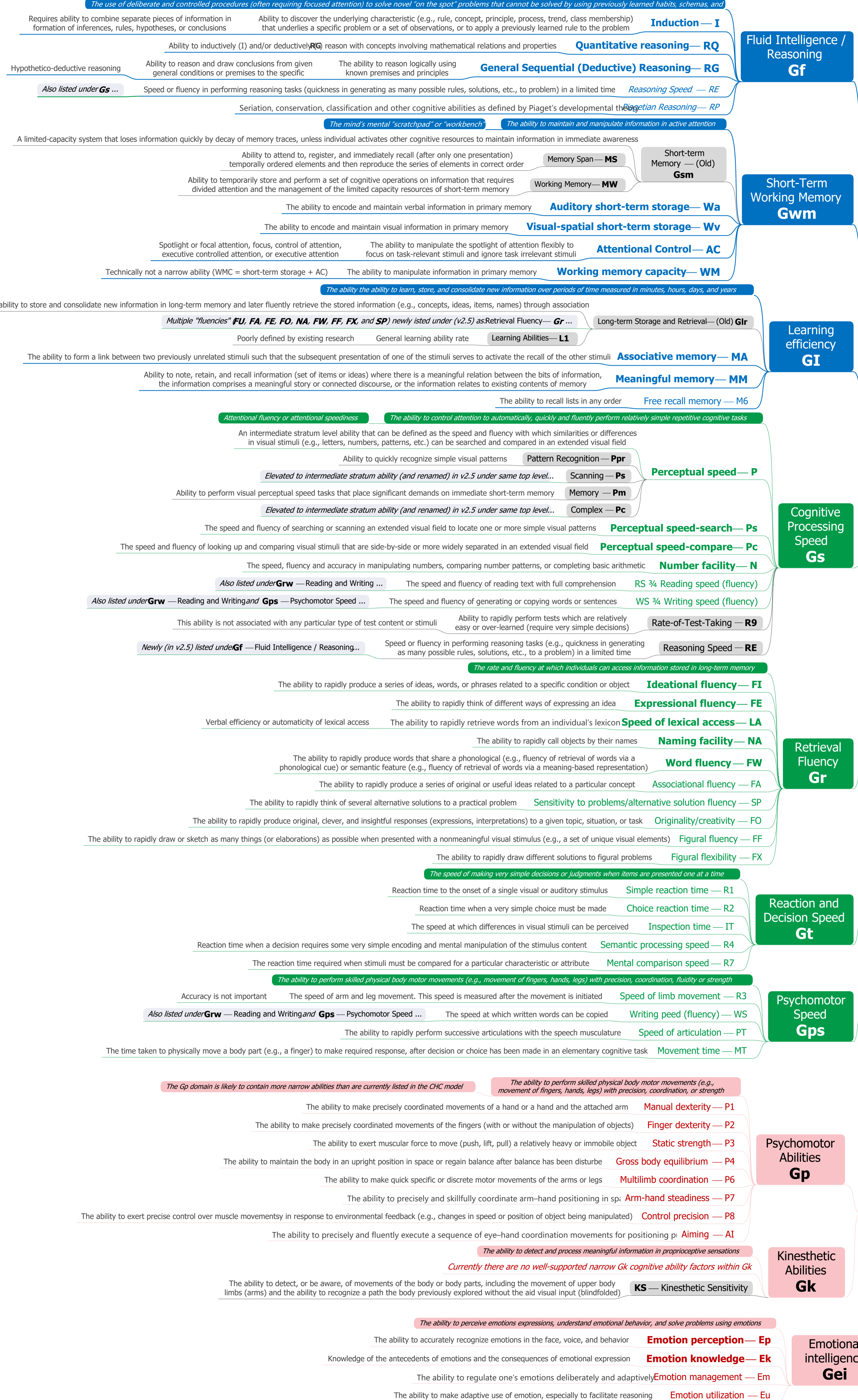
Iterate this process thousands of times until a viable architecture is discovered

Google, "Using Machine Learning to Explore Neural Network Architecture"

AutoML is a recent method that has grown out of a long history of a broad class of neuro-evolutionary methods

# **Appendix H: Cattell-Horn-Carroll (CHC) Taxonomy of Cognitive Abilities**

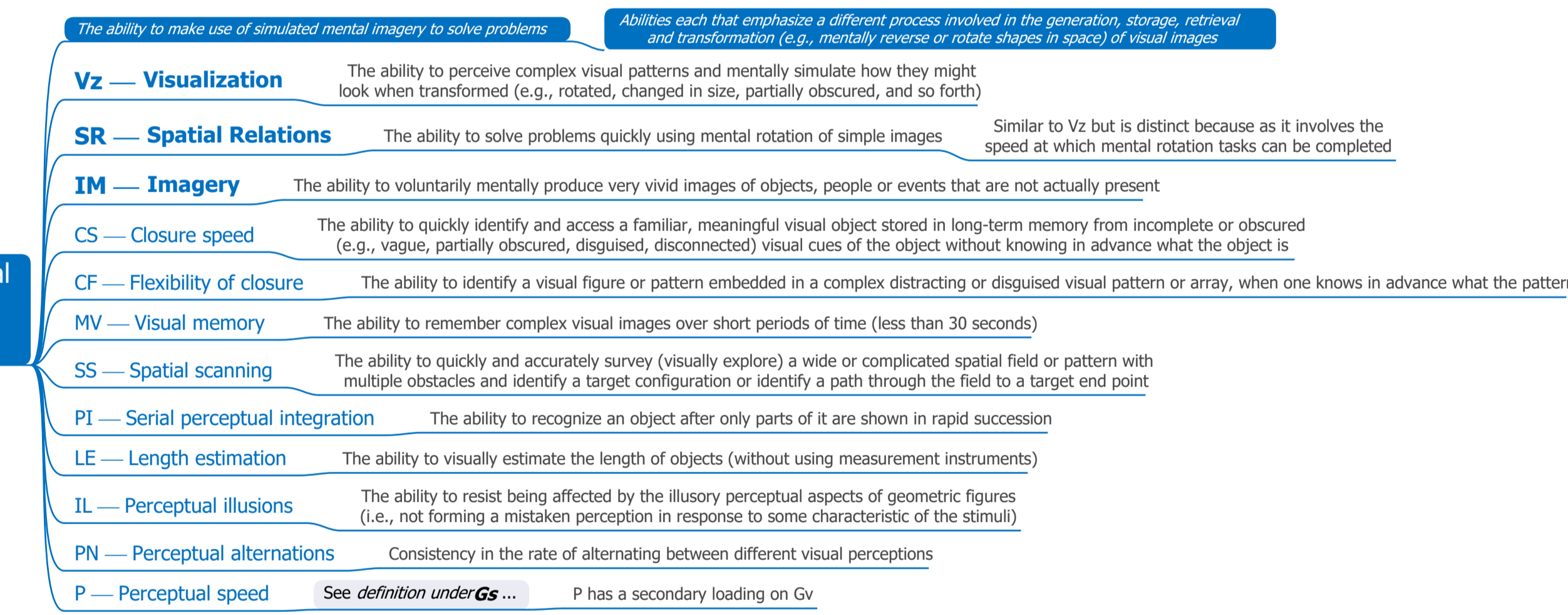
---



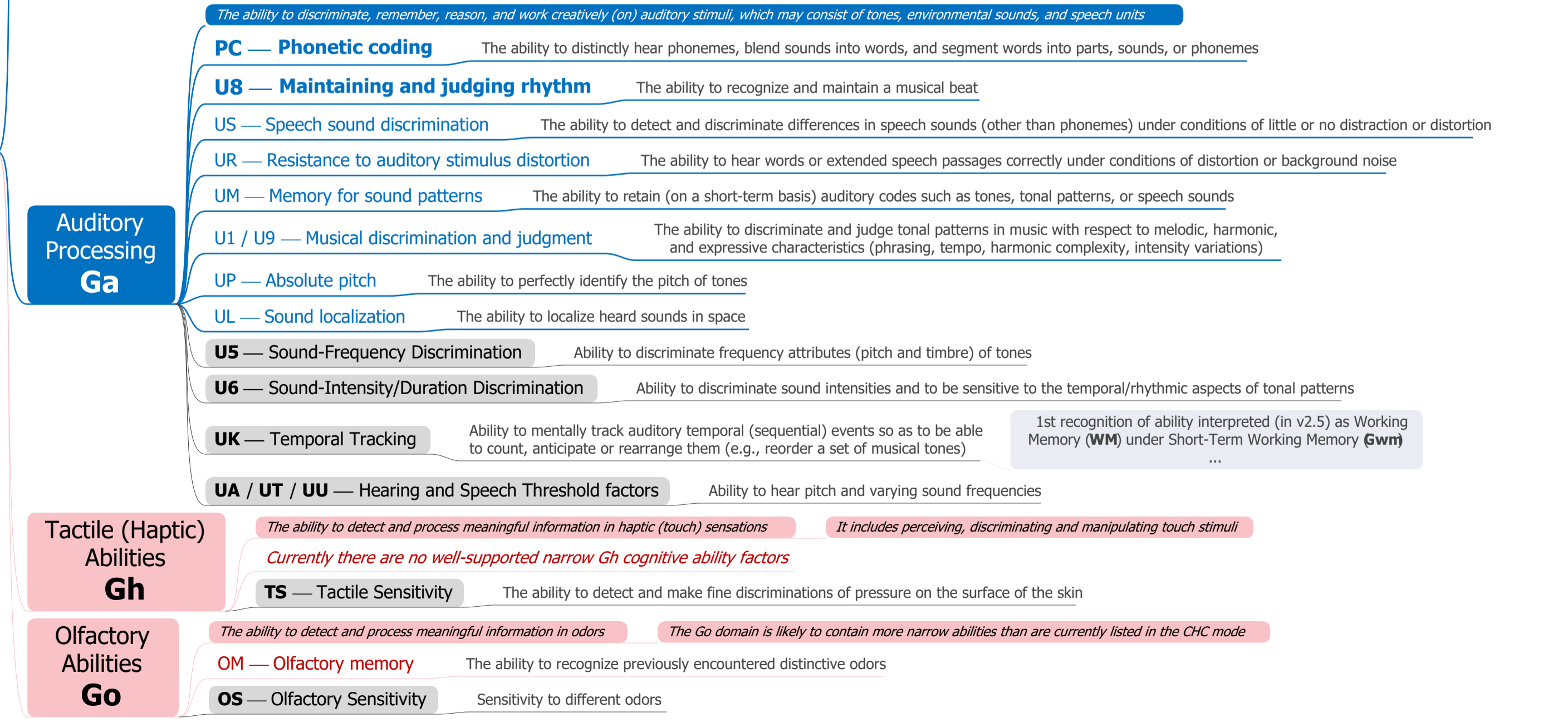
**Acquired Knowledge Systems**



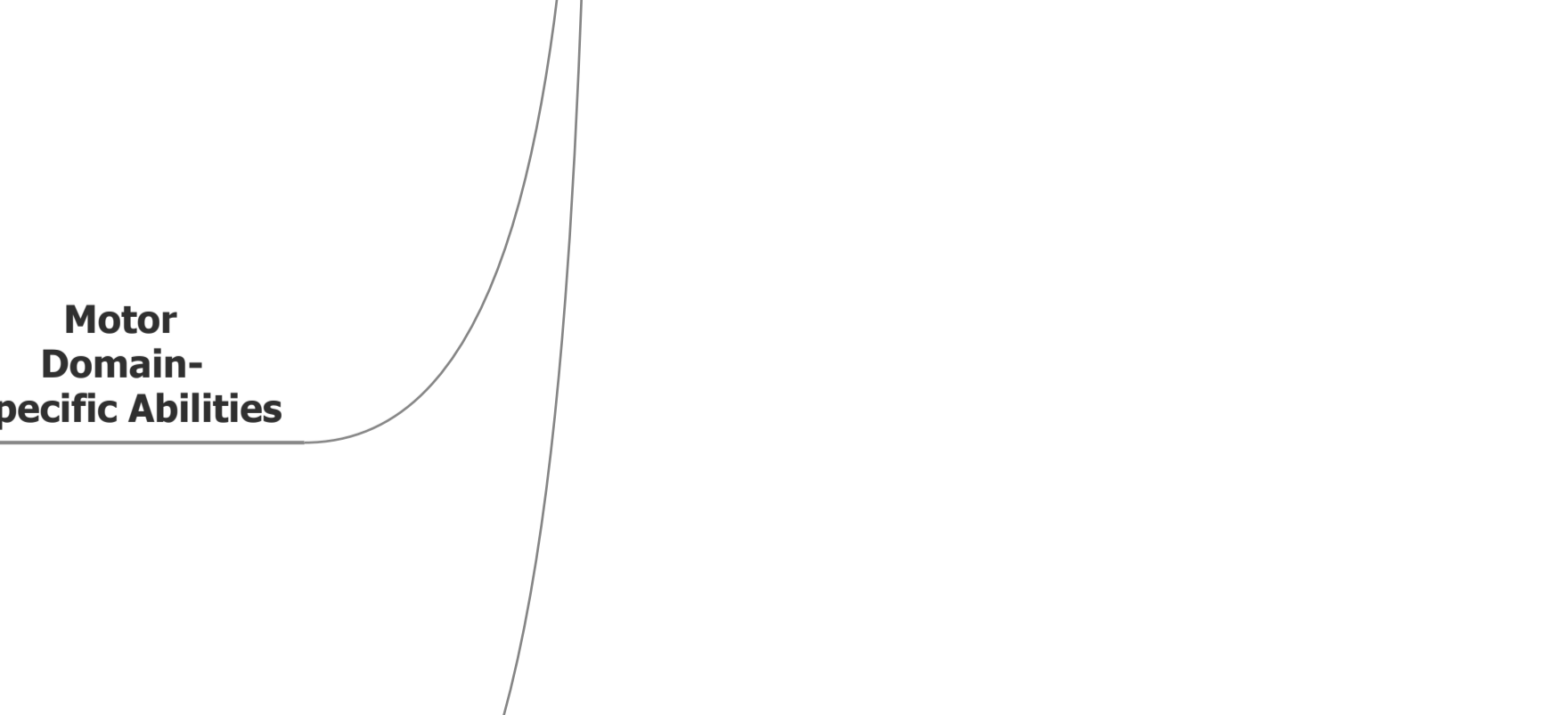
**Visual-Spatial Abilities (Gv)**



**Sensory Domain-Specific Abilities**



**Motor Domain-Specific Abilities**



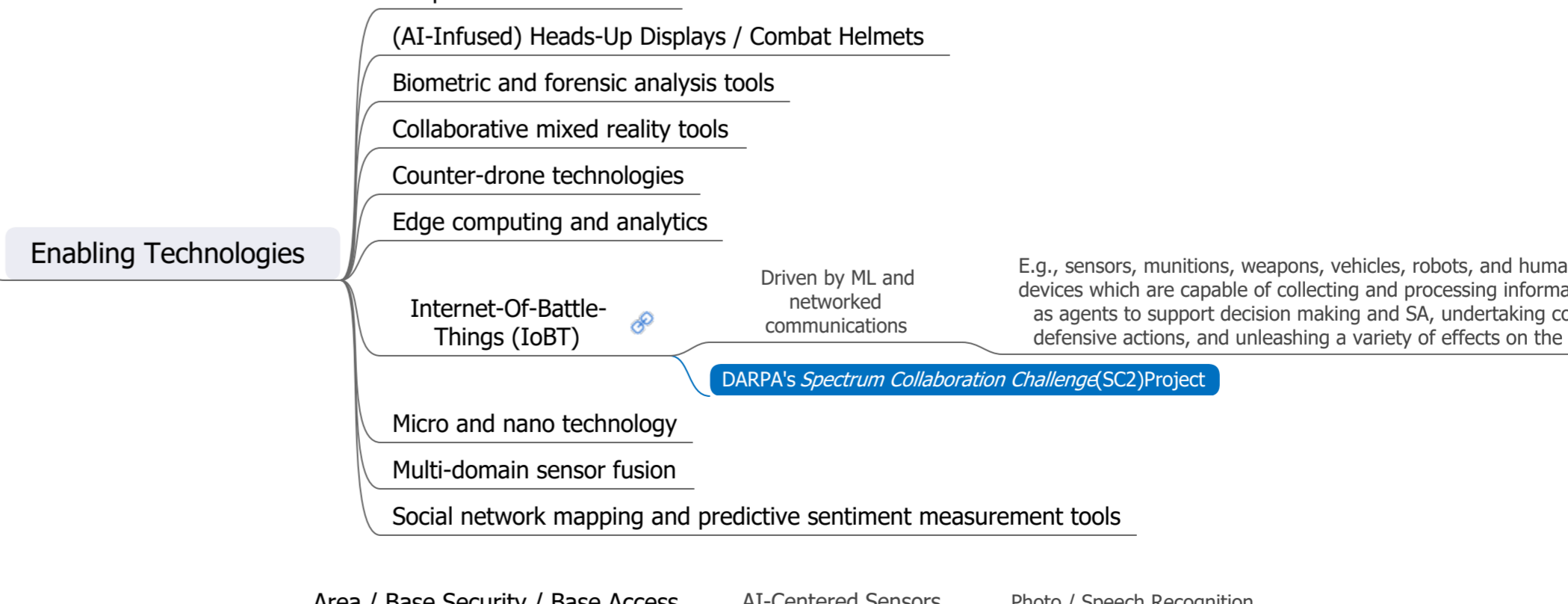
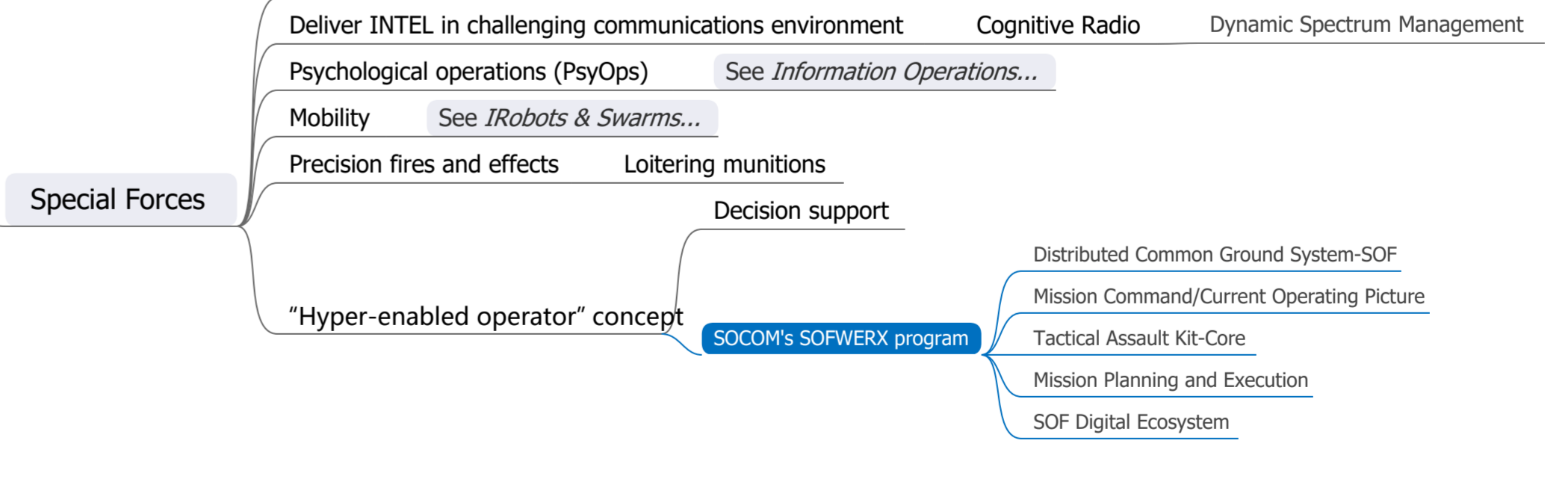
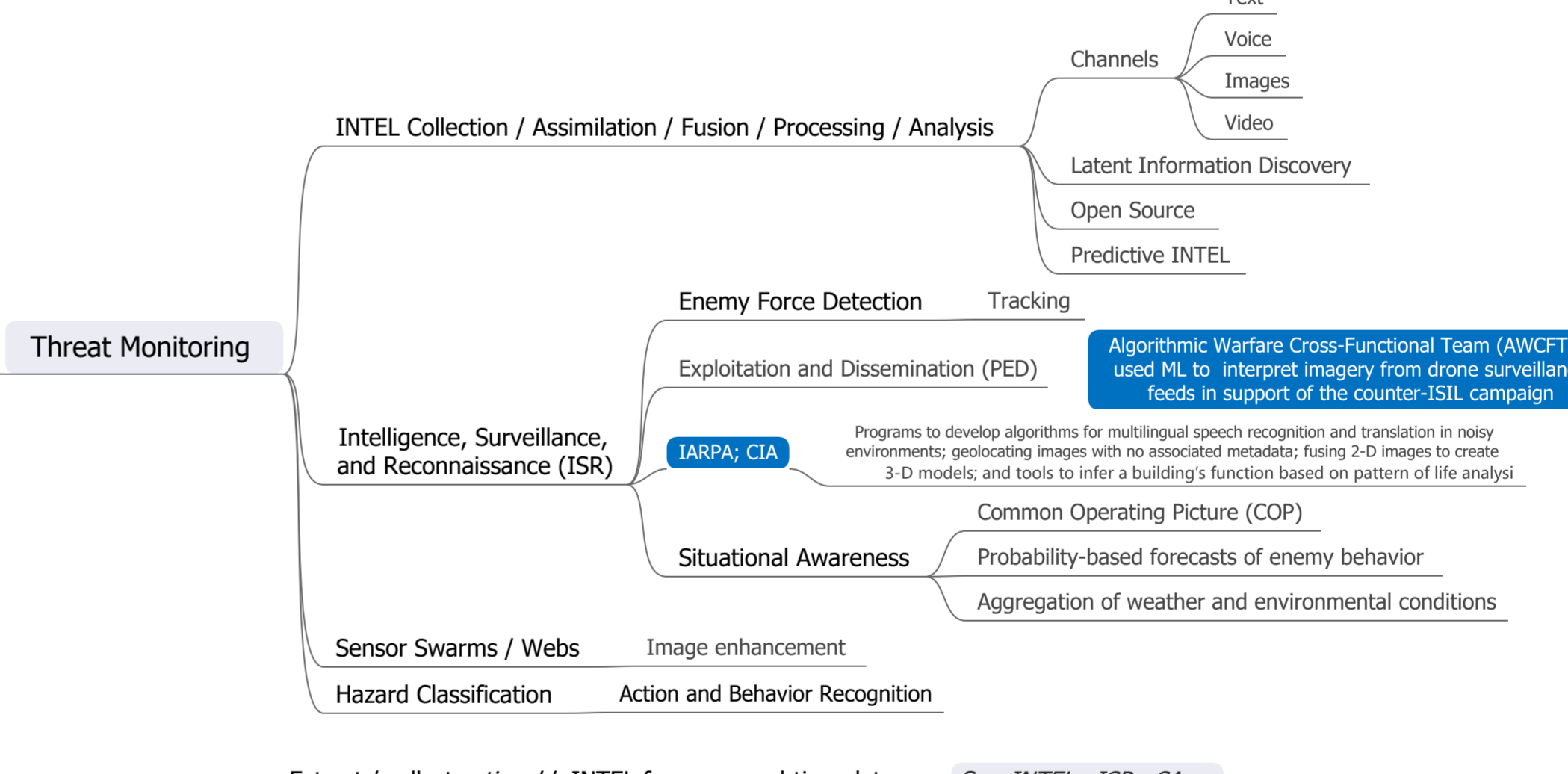
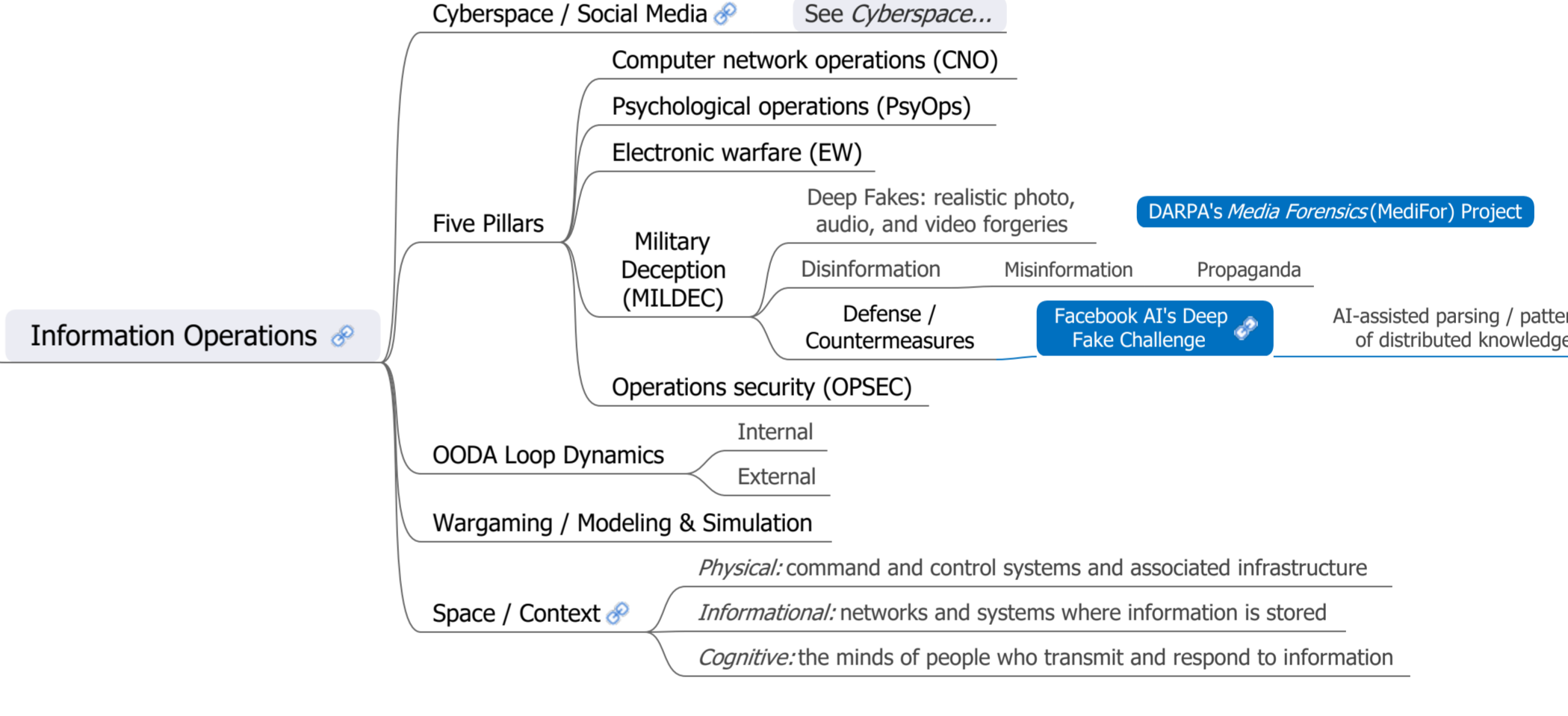
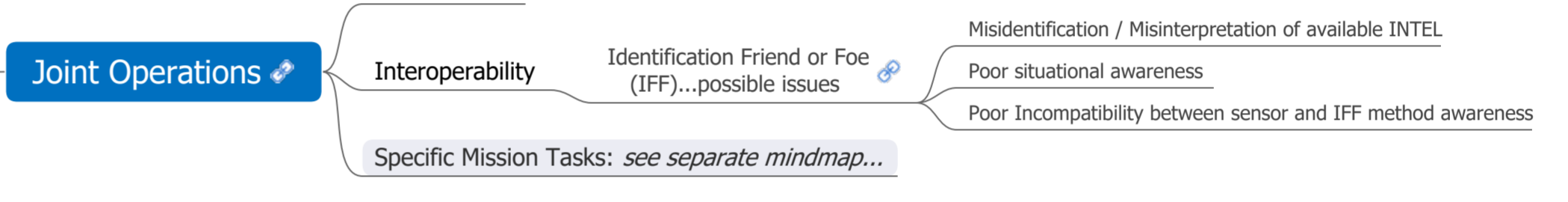
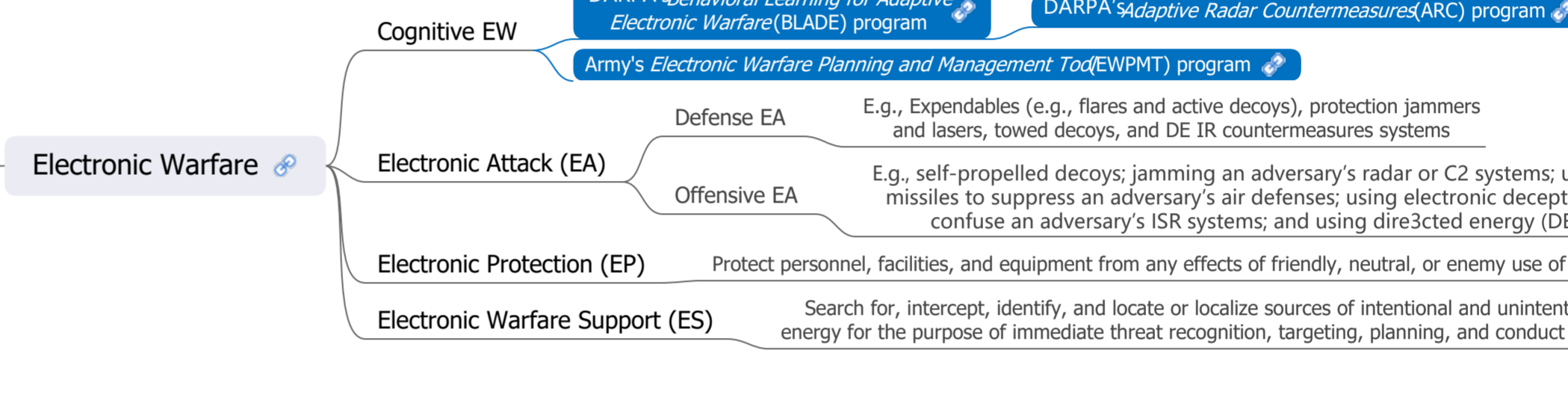
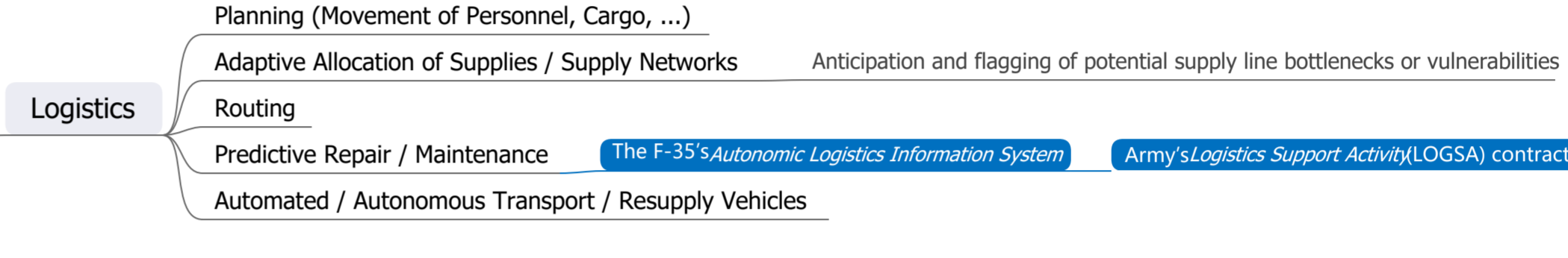
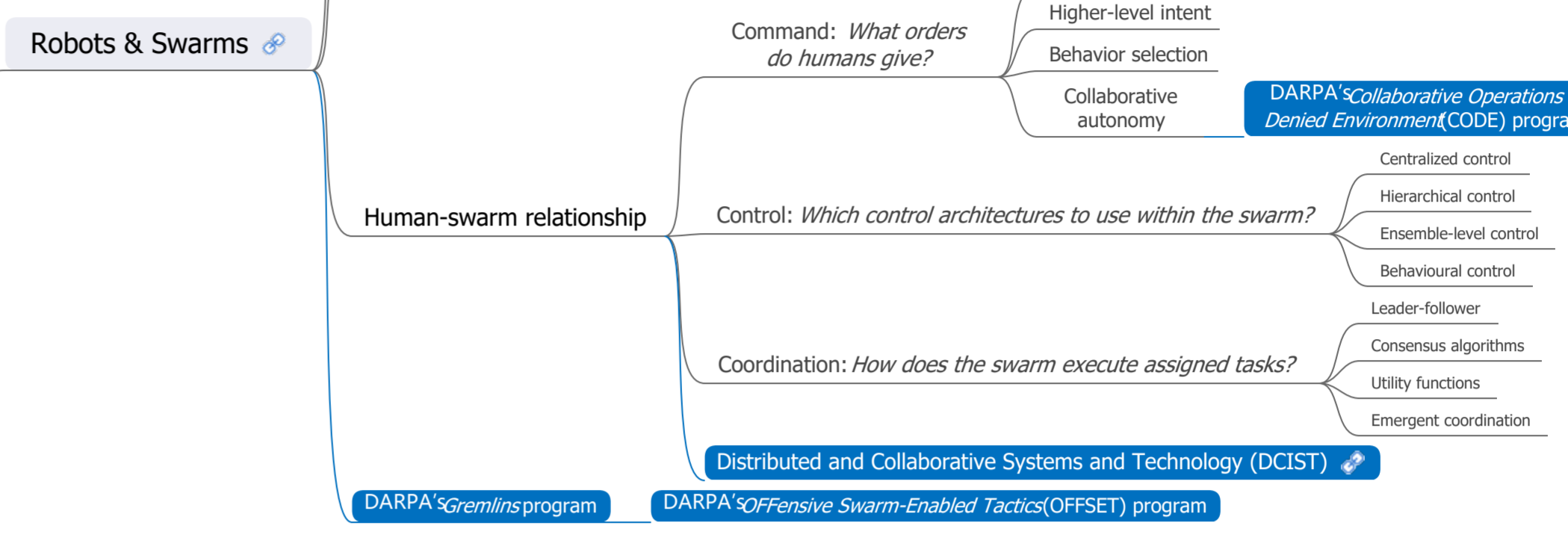
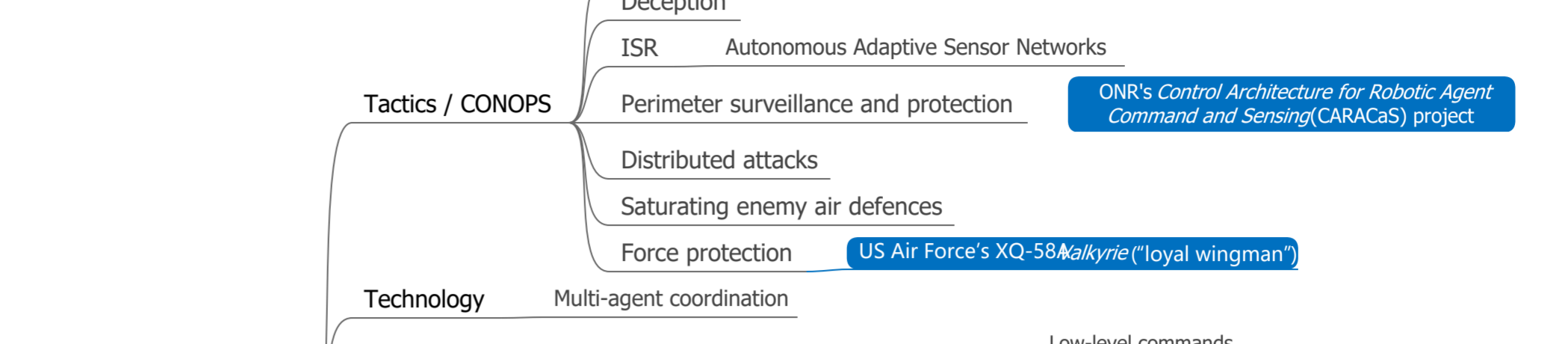
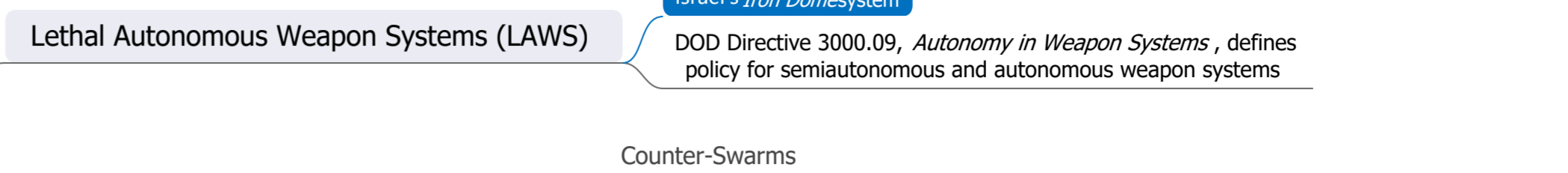
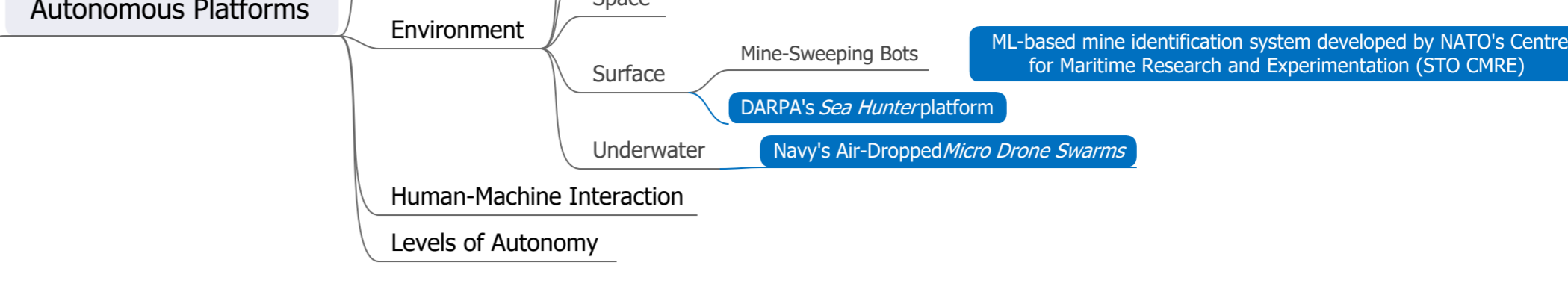
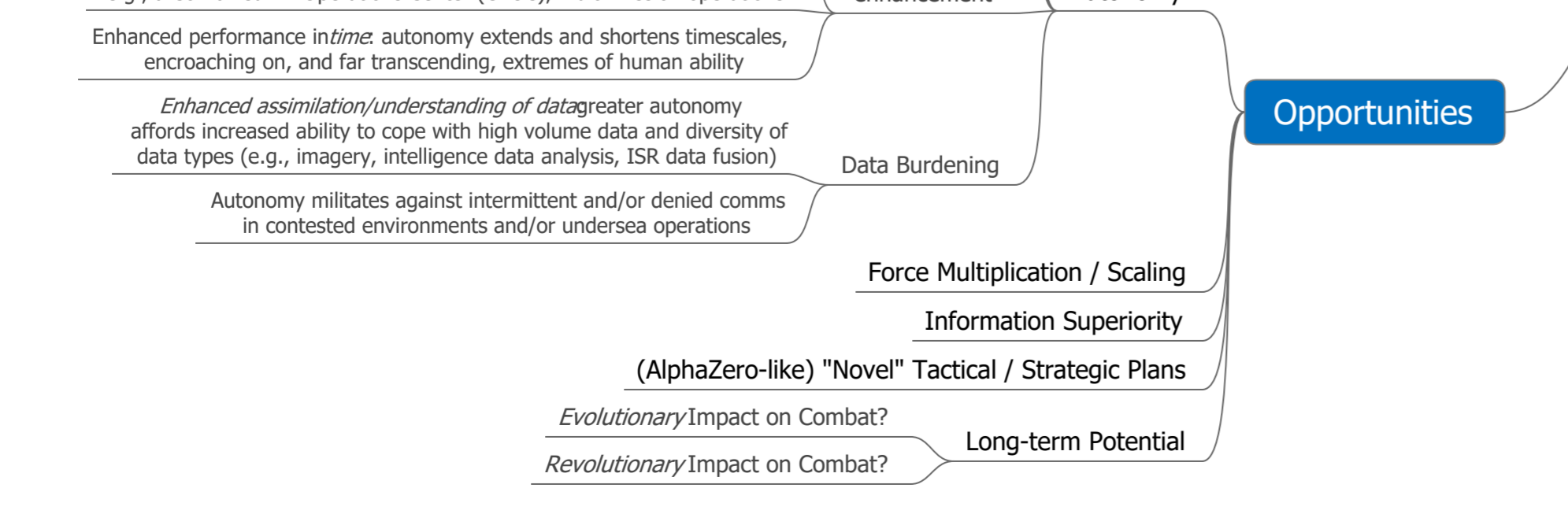
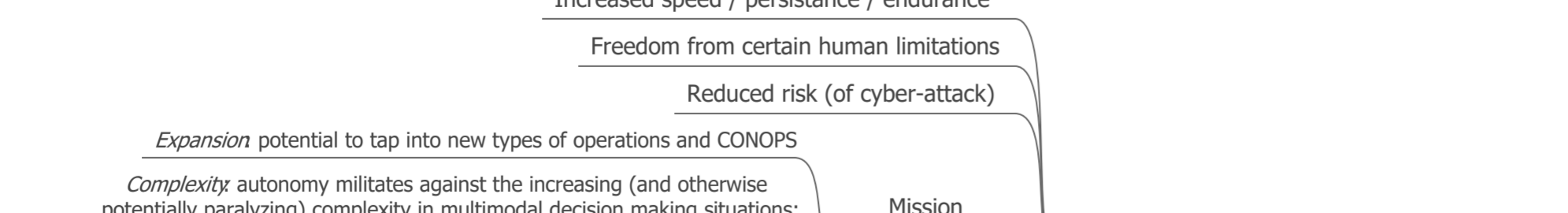
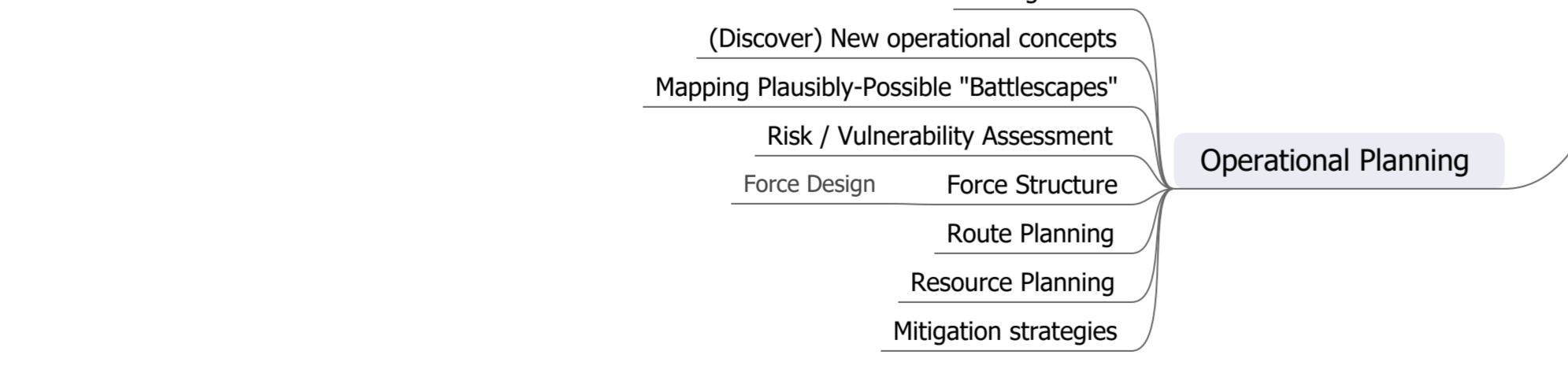
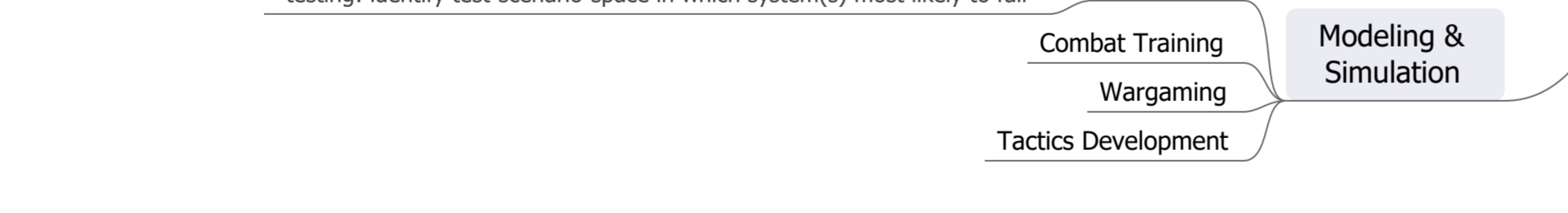
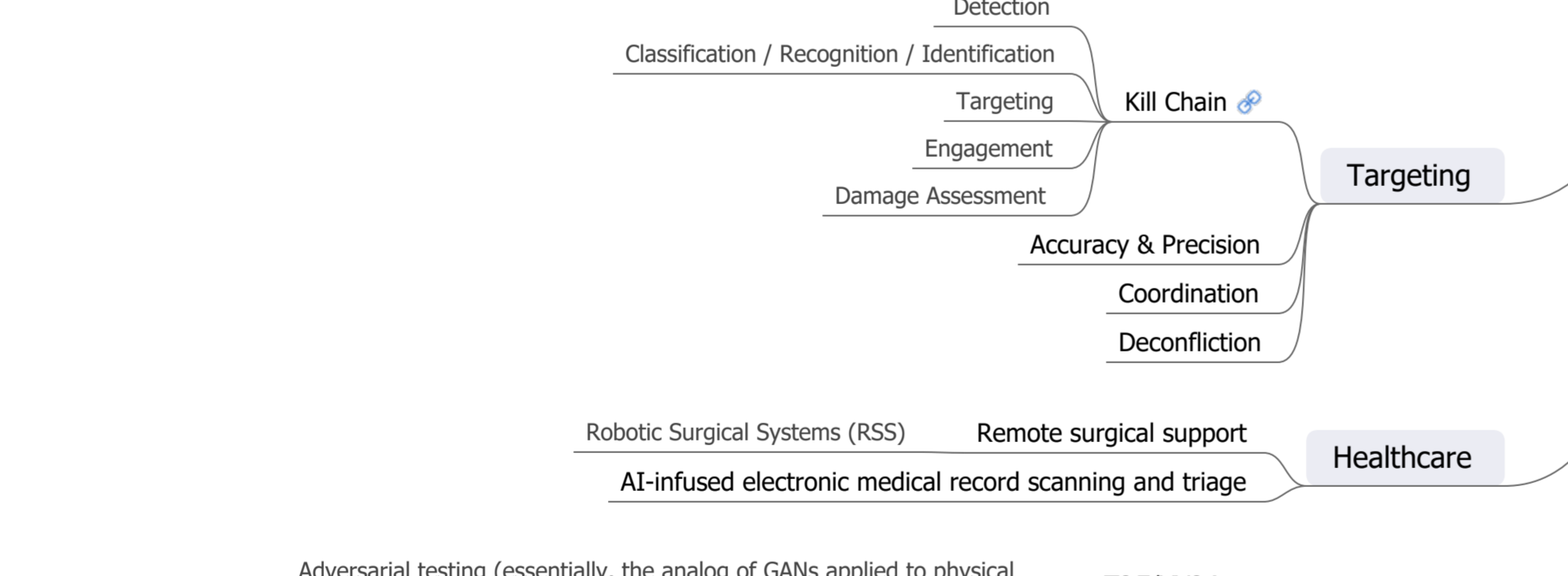
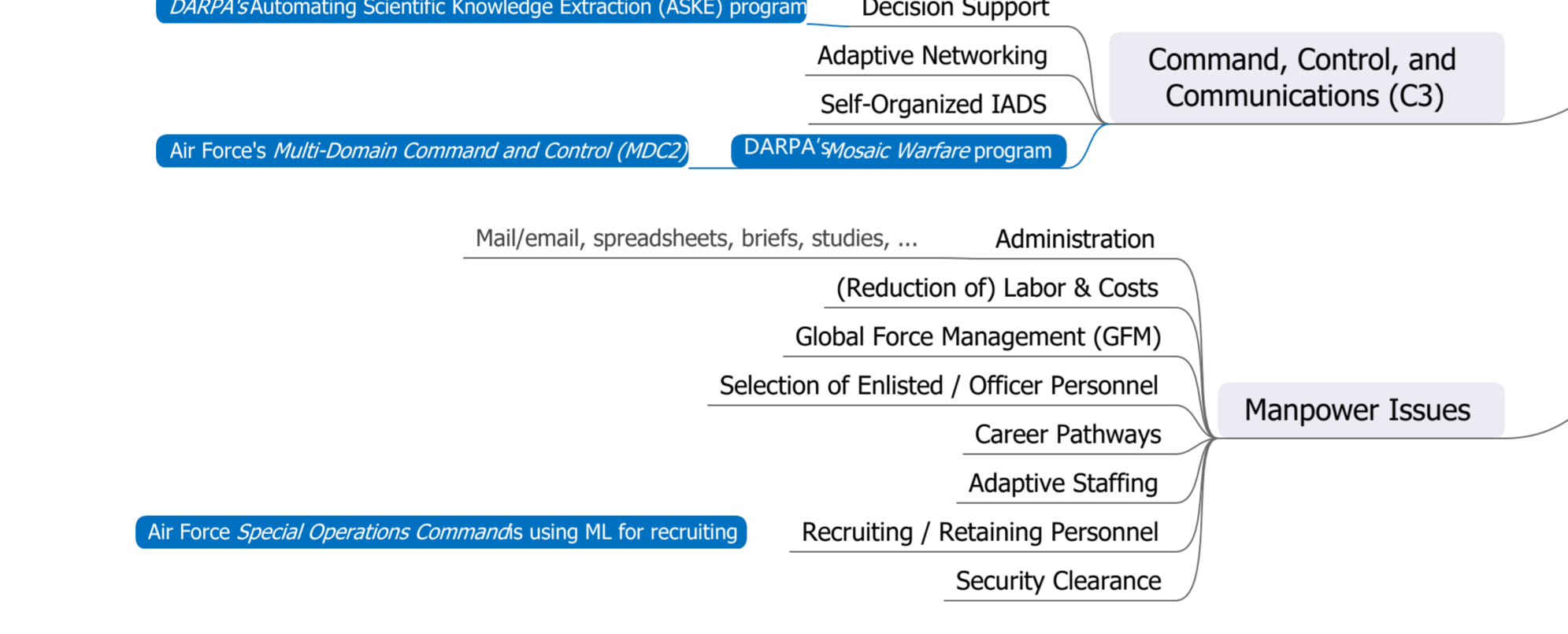
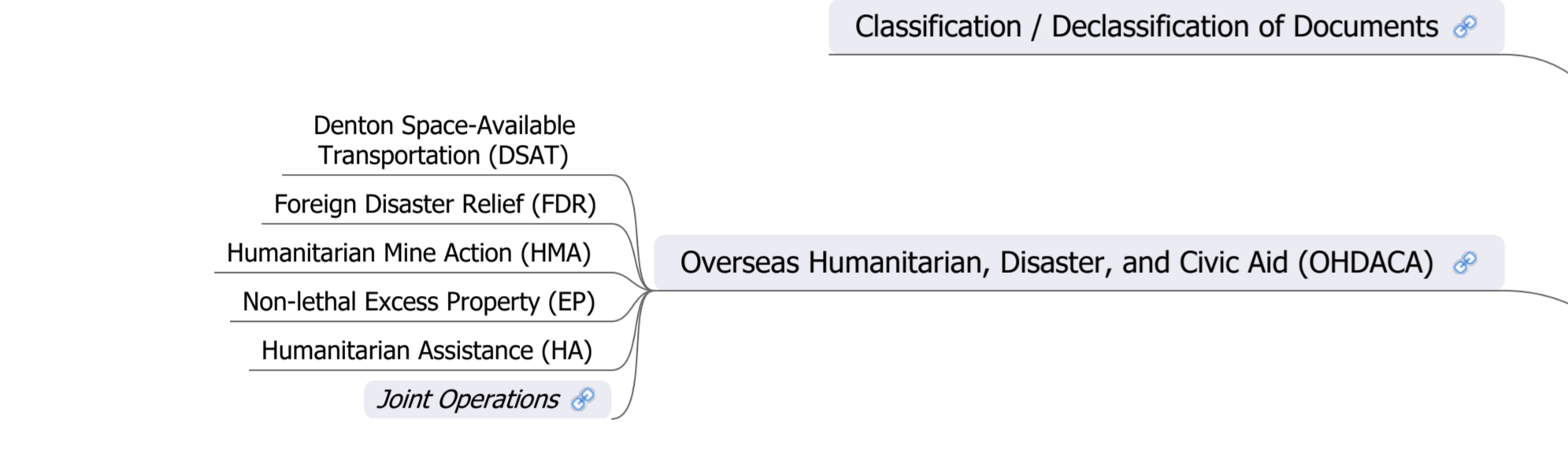
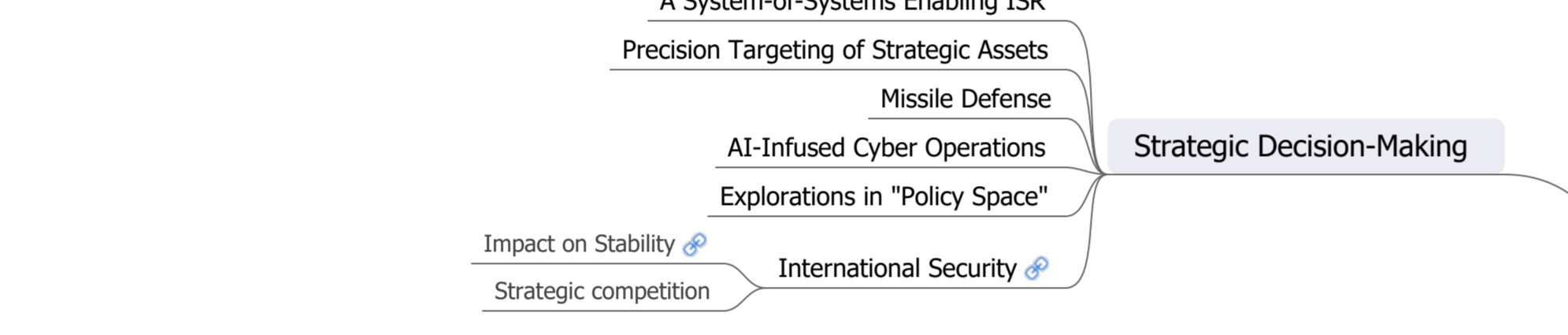
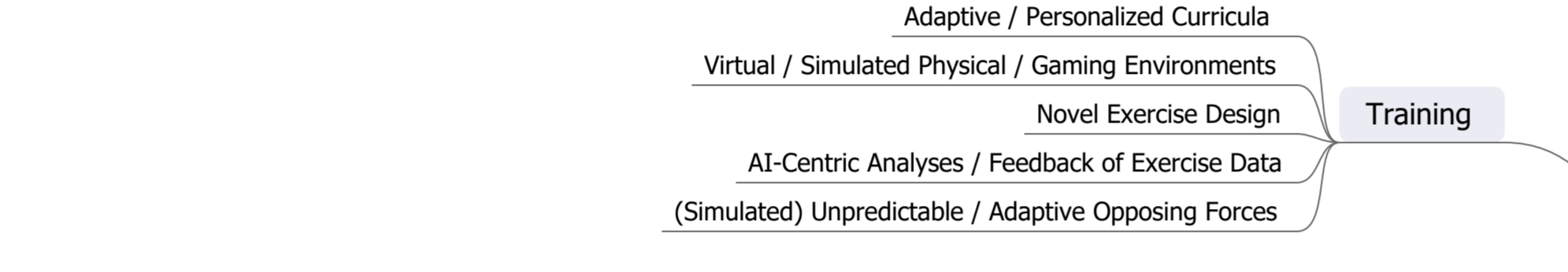
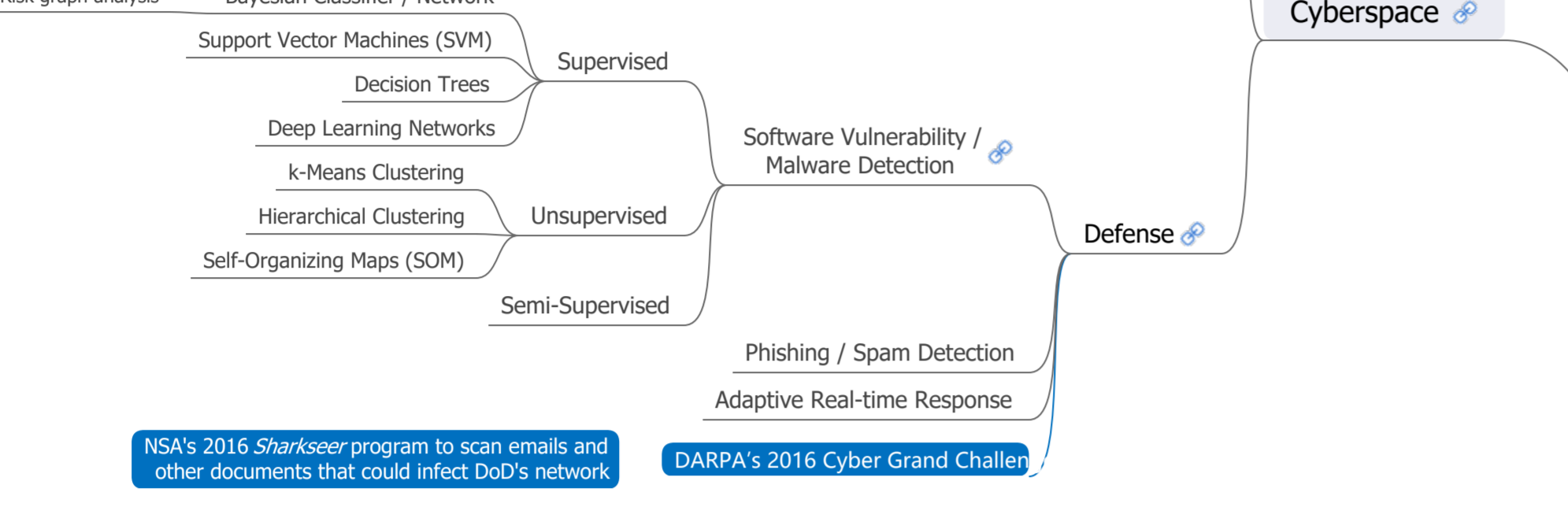
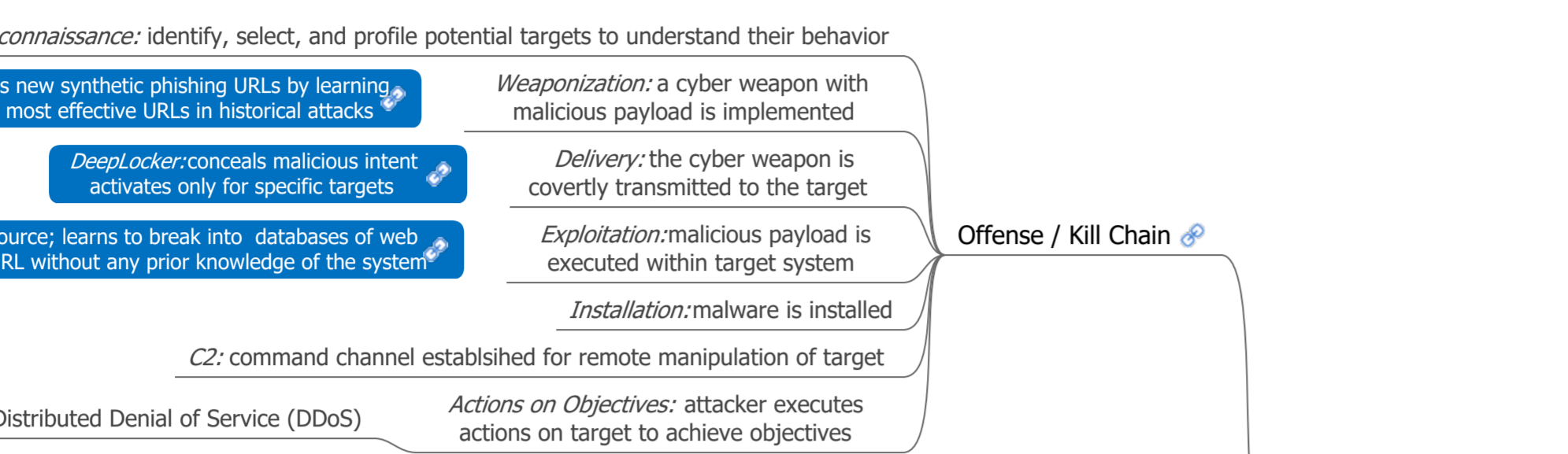
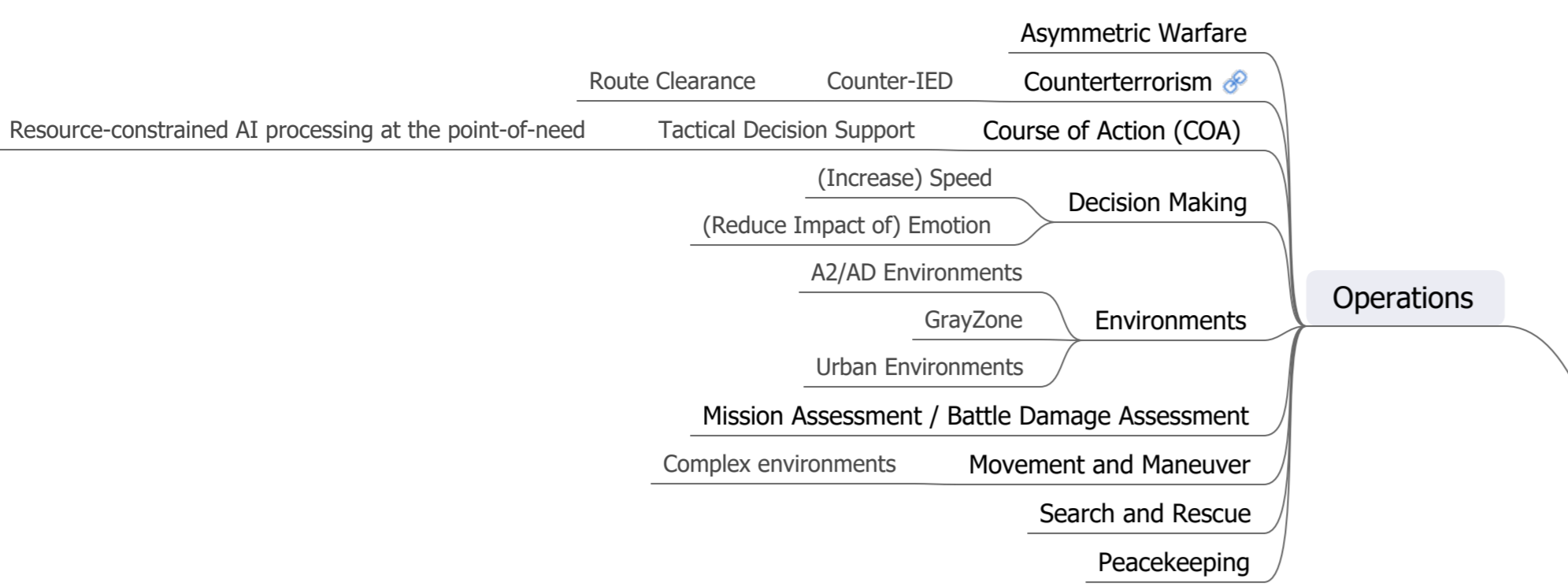
**Emotional Domain**



# Appendix I: Mindmap of Possible Military Applications of AI

---

# Possible Military Applications of AI





# Figures

---

Figure 1. Visual schematic of the major sections of this paper.....	3
Figure 2. A mindmap/timelines of significant JAIC-related milestones, 2017–July 2020 .....	7
Figure 3. A mindmap/timelines of DARPA AI-related program announcements, 2017– July 2020 .....	10
Figure 4. Timeline of milestones in the development of neural networks and deep learning.....	24
Figure 5. Schematic illustrations of neural network designs .....	26
Figure 6. A schematic illustration of the "Five Tribes" of AI.....	31
Figure 7. A visual taxonomy of ML methods and functions (top-level view only).....	33
Figure 8. Mindmap of milestone developments of NN designs and architectures.....	34
Figure 9. A schematic illustration of a typical AI/ML development pipeline .....	37
Figure 10. Google Trends statistics for five AI-related key phrases .....	44
Figure 11. Number of AI-related papers posted to ArXiv.CS between 2009 and 2020.....	45
Figure 12. Number of papers posted to selected AI/ML-related branches of the ArXiv.CS.....	46
Figure 13. Number of AI-related papers posted to ArXiv.CS that satisfy specific search phrases.....	47
Figure 14. A timeline of notable AI achievements, mid-year 2017 through end of 2018.....	52
Figure 15. A timeline of notable AI achievements, 2019 through mid-year 2020 .....	53
Figure 16. Summary of the CHC cognitive abilities framework.....	76
Figure 17. OODA loop and elements pertaining to AI and ML.....	79
Figure 18. OODA loop–based autonomy taxonomy .....	81
Figure 19. Toward enfolding AI/ML, the OODA loop, and the CHC cognitive framework ...	82
Figure 20. A concept for a CHC-mediated bridge between military operations and AI methodology .....	83
Figure 21. Notional examples of CHC-based psychometric profiles of "general intelligence" ..	87
Figure 22. List of shape and color codes used to highlight mindmap entries.....	96
Figure 23. "AI with AI" podcast corpus summary .....	96
Figure 24. SOTA + Milestone Achievements + Innovative Concepts.....	98
Figure 25. SOTA + Milestone Achievements + Innovative Concepts: <i>Counts</i> .....	99
Figure 26. Limitations, Vulnerabilities, and "Anti AI" Backlashes.....	100
Figure 27. Limitations, Vulnerabilities, and "Anti AI" Backlashes: <i>Counts</i> .....	101

[This page intentionally left blank]

# Abbreviations

---

ACE	Air Combat Evolution
AGI	Artificial General Intelligence
AI	Artificial Intelligence
ANN	Artificial Neural Network
AutoML	Automated Machine Learning
BM	Boltzmann Machine
CDC	Centers for Disease Control and Prevention
CNN	Convolutional Neural Network
CORD-19	COVID-19 Open Research Dataset
CS	Computer Science
CVPR	Conference on Computer Vision and Pattern Recognition
DARPA	Defense Advanced Research Projects Agency
DIB	Defense Innovation Board
DL	Deep Learning
DLNN	Deep Learning Neural Network
DOD	Department of Defense
DODIG	DOD Inspector General
EA	Evolutionary Algorithm
ES	Expert Systems
FLOP	Floating-Point Operation
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
HN	Hopfield Network
IR	Information Retrieval
JAIC	Joint Artificial Intelligence Center
LSTM	Long Short-Term Memory
ML	Machine Learning
NAS	Neural Architecture Search
NDAA	National Defense Authorization Act
NESTA	National Endowment for Science Technology and the Arts
NIH	National Institutes of Health
NLP	Natural Language Processing
NMI	National Mission Initiative
NN	Neural Network
NSCAI	National Security Commission on Artificial Intelligence

OODA	Observe-Orient-Decide-Act
PED	Processing, Exploitation, and Dissemination
RDSC	Roche Data Science Coalition
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SOTA	State of the Art
SVM	Support Vector Machine
T&E	Testing and Evaluation
TSoM	The Society of Mind
TST	Three-Stratum Theory
VV&A	Verification, Validation, and Accreditation
WAMI	Wide-Area Motion Imagery
WHO	World Health Organization
XAI	Explainable Artificial Intelligence

# References: Surveys of AI/ML Research

---

- Abdallah, Tarek and Beatriz de La Iglesia. 2015. *Survey on Feature Selection*. <https://arxiv.org/pdf/1510.02892>.
- Abiodun, Oludare Issac, et al. 2018. *State-of-the-art in artificial neural network applications: A survey*. <https://www.sciencedirect.com/science/article/pii/S2405844018332067>.
- Ackerman, Joshua and George Cybenko. 2020. *A Survey of Neural Networks and Formal Languages*. <https://arxiv.org/pdf/2006.01338>.
- Alom, Zahangir, et al. 2018. *The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches*. <https://arxiv.org/ftp/arxiv/papers/1803/1803.01164.pdf>.
- Alom, Zahangir, et al. 2019. "A State-of-the-Art Survey on Deep Learning Theory and Architectures," *Electronics*, Vol. 8. <https://www.mdpi.com/2079-9292/8/3/292/pdf-vor>.
- Alsharif, Mohammed, et al. 2020. *Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment: Taxonomies and Research Trends*. <https://www.mdpi.com/2073-8994/12/1/88>.
- Arora, Saurabh and Prashant Doshi. 2018. *A Brief Survey of Deep Reinforcement Learning*. <https://arxiv.org/pdf/1806.06877>.
- Arrieta, A., et al. 2020. *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*. <https://arxiv.org/pdf/1910.10045>.
- Arthur, Aubret, et al. 2017. *A survey on intrinsic motivation in reinforcement learning*. <https://arxiv.org/pdf/1908.06976>.
- Arulkumaran, Kai, et al. 2017. *A Brief Survey of Deep Reinforcement Learning*. <https://arxiv.org/pdf/1708.05866>.
- Batmaz, Z., et al. 2019. *A review on deep learning for recommender systems: challenges and remedies*. [https://daiwk.github.io/assets/Batmaz2018\\_Article\\_AReviewOnDeepLearningForRecomm.pdf](https://daiwk.github.io/assets/Batmaz2018_Article_AReviewOnDeepLearningForRecomm.pdf).
- Battaglia, Peter, et al. 2018. *Relational inductive biases, deep learning, and graph networks*. <https://arxiv.org/pdf/1806.01261>.
- Bekker, Jessa and Jesse Davis. 2020. *Learning from positive and unlabeled data: a survey*. <https://arxiv.org/pdf/1811.04820>.
- Belinkov, Yonatan and James Glass. 2019. *Analysis Methods in Neural Language Processing: A Survey*. <https://arxiv.org/pdf/1812.08951>.
- Bendre, Nihar, et al. 2020. *Learning from Few Samples: A Survey*. <https://export.arxiv.org/pdf/2007.15484>.
- Besold, Tarek, et al. 2017. *Neural-Symbolic Learning and Reasoning: A Survey and Interpretation*. <https://arxiv.org/pdf/1711.03902>.
- Bullock, Joseph, et al. 2020. *Mapping the Landscape of Artificial Intelligence Applications against COVID-19*. <https://arxiv.org/pdf/2003.11336>.

- Bekker, Jessa and Jesse Davis. 2020. *Learning from positive and unlabeled data: a survey*. <https://arxiv.org/pdf/1811.04820>.
- Chen, Jianguo. 2020. *A Survey on Applications of Artificial Intelligence in Fighting Against COVID-19*. <https://arxiv.org/pdf/2007.02202>.
- Chen, Xieling. 2020. *Topics and trends in artificial intelligence assisted human brain research*. *PLOSOne*. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231192>.
- Chakraborty, Anirban, et al. 2018. *Adversarial Attacks and Defences: A Survey*. <https://arxiv.org/pdf/1810.00069>.
- Chari, Shruthi. 2020. *Foundations of Explainable Knowledge-Enabled Systems*. <https://arxiv.org/pdf/2003.07520>.
- Creswell, Antonia, et al. 2020. *Generative Adversarial Networks: An Overview*. <https://arxiv.org/pdf/1710.07035>.
- Chen, Xiaoxue, et al. 2020. *Text Recognition in the Wild: A Survey*. <https://arxiv.org/pdf/2005.03492>.
- Coppola, Mario, et al. 2020. *A Survey on Swarming With Micro Air Vehicles: Fundamental Challenges and Constraints*. <https://www.frontiersin.org/articles/10.3389/frobt.2020.00018/full>.
- Doersch, Carl. 2016. *Tutorial on Variational Autoencoders*. <https://arxiv.org/pdf/1606.05908.pdf>.
- Das, Arun and Paul Rad. 2020. *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*. <https://arxiv.org/pdf/2006.11371>.
- Elksen, Thomas, et al. 2019. *Neural Architecture Search: A Survey*. <https://arxiv.org/pdf/1808.05377>.
- Elton, Daniel C. 2020. *Self-explaining AI as an alternative to interpretable AI*. <https://arxiv.org/pdf/2002.05149>.
- Engelen, Jesper and Holger Hoos. 2019. *A survey on semi-supervised learning*. <https://link.springer.com/article/10.1007/s10994-019-05855-6>.
- Fawaz, Hassan, et al. 2019. *Deep learning for time series classification: a review*. <https://arxiv.org/pdf/1809.04356>.
- Galassi, Andrea, et al. 2019. *Attention in Natural Language Processing*. <https://arxiv.org/pdf/1902.02181v2>.
- Gambella, Claudio, et al. 2019. *Optimization Models for Machine Learning: A Survey*. <https://arxiv.org/pdf/1901.05331>.
- Garmacea, Cristina and Qiaozhu Mei. 2020. *Neural Language Generation: Formulation, Methods, and Evaluation*. <https://arxiv.org/pdf/2007.15780>.
- Ghods, Alireza and Diane Cook. 2019. *A Survey of Techniques All Classifiers Can Learn from Deep Networks: Models, Optimizations, and Regularization*. <https://arxiv.org/pdf/1909.04791>.
- Gilber, Daniel, et al. 2020. *The rise of machine learning for detection and classification of malware: Research developments, trends and challenges*. <https://www.sciencedirect.com/science/article/pii/S1084804519303868>.
- Gou, Jianping, et al. 2020. *Knowledge Distillation: A Survey*. <https://arxiv.org/pdf/2006.05525>.

- Grando, Felipe. 2018. *Machine Learning in Network Centrality Measures: Tutorial and Outlook*. <https://dl.acm.org/doi/10.1145/3237192>.
- Gui, Jie, et al. 2020. *A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications*. <https://arxiv.org/pdf/2001.06937>.
- Guidotti, Riccardo, et al. 2018. *A Survey Of Methods For Explaining Black Box Models*. <https://dl.acm.org/doi/10.1145/3236009>.
- Gupta, Pranshu. 2019. *Algorithms Inspired by Nature: A Survey*. <https://arxiv.org/pdf/1903.01893>.
- Hartley, Matthew and Tjelvar Olsson. 2020. *dtoolAI: Reproducibility for Deep Learning*. <https://www.sciencedirect.com/science/article/pii/S2666389920300933>.
- He, Xin, et al. 2019. *AutoML: A Survey of the State-of-the-Art*. <https://arxiv.org/pdf/1908.00709>.
- He, Zhiyuan, et al. 2019. *Gradient Boosting Machine: A Survey*. <https://arxiv.org/pdf/1908.06951>.
- Hogan, Aidan, et al. 2020. *Knowledge Graphs*. <https://arxiv.org/pdf/2003.02320>.
- Hospedales, Timothy, et al. 2020. *Meta-Learning in Neural Networks: A Survey*. <https://arxiv.org/pdf/2004.05439>.
- Hu, Dichao. 2019. *An Introductory Survey on Attention Mechanisms in NLP Problems*. <https://arxiv.org/pdf/1811.05544>.
- Hug, Timothy, et al. 2020. *Adversarial Attacks and Defense on Texts: A Survey*. <https://arxiv.org/pdf/2005.14108>.
- Huang, Xiaowei, et al. 2018. *A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability*. <https://arxiv.org/pdf/1812.08342>.
- Humbatova, Nargiz, et al. 2019. *Taxonomy of Real Faults in Deep Learning Systems*. <https://arxiv.org/pdf/1910.11015>.
- Islam, Md Johirul, et al. 2019. *A Comprehensive Study on Deep Learning Bug Characteristics*. <https://arxiv.org/pdf/1906.01388>.
- Jabbar, Abdul, et al. 2020. *A Survey on Generative Adversarial Networks: Variants, Applications, and Training*. <https://arxiv.org/pdf/2006.05132>.
- Jauhiainen, Tommi. 2019. *Automatic Language Identification in Texts: A Survey*. <https://arxiv.org/pdf/1804.08186>.
- Ji, Shaoxiong, et al. 2020. *A Survey on Knowledge Graphs: Representation, Acquisition and Applications*. <https://arxiv.org/pdf/2002.00388>.
- Jiménez-Luna, José, et al. 2020. *Drug discovery with explainable artificial intelligence*. <https://arxiv.org/pdf/2007.00523>.
- Jin, Yilun, et al. 2020. *A Survey towards Federated Semi-supervised Learning*. <https://arxiv.org/pdf/2002.11545v1>.
- Kaufmann, Elia, et al. 2020. *Deep Drone Acrobatics*. <https://arxiv.org/pdf/2006.05768>.

- Kazemi, Seyed Mehran, et al. 2019. *Relational Representation Learning for Dynamic (Knowledge) Graphs: A Survey*. <https://arxiv.org/pdf/1905.11485>.
- Kinderkhedra, Mital. 2019. *Learning Representations of Graph Data -- A Survey*. <https://arxiv.org/pdf/1906.02989>.
- Koturwar, Praful, et al. 2020. *A Survey of Classification Techniques in the Area of Big Data*. <https://arxiv.org/pdf/1503.07477>.
- Kowsari, Kamran, et al. 2020. *Text Classification Algorithms: A Survey*. <https://arxiv.org/pdf/1904.08067>.
- Khurana, Diksha, et al. 2017. *Natural Language Processing: State of The Art, Current Trends and Challenges*. <https://arxiv.org/pdf/1708.05148>.
- Kumar, Atul and Sameep Mehta. 2017. *A Survey on Resilient Machine Learning*. <https://arxiv.org/pdf/1707.03184>.
- Lamb, Luis, et al. 2020. *Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective*. <https://arxiv.org/pdf/2003.00330>.
- Lathuilière, Stéphane, et al. 2018. *A Comprehensive Analysis of Deep Regression*. <https://arxiv.org/pdf/1803.08450>.
- LeCun, Yann, 2015. *Deep Learning*. *Nature*. Vol. 521. <https://www.nature.com/articles/nature14539>.
- Labach, Alex, et al. 2019. *Survey of Dropout Methods for Deep Neural Networks*. <https://arxiv.org/pdf/1904.13310>.
- Lehman, Joel, et al. 2018. *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*. <https://arxiv.org/pdf/1803.03453>.
- Lim, Bryan and Stefan Zohren. 2020. *Time Series Forecasting With Deep Learning: A Survey*. <https://arxiv.org/pdf/2004.13408>.
- Li, Jing, et al. 2018. *A Survey on Deep Learning for Named Entity Recognition*. <https://arxiv.org/pdf/1812.09449>.
- Li, Mingzhen, et al. 2020. *The Deep Learning Compiler: A Comprehensive Survey*. <https://arxiv.org/pdf/2002.03794>.
- Li, Yuxi. 2017. *Deep Reinforcement Learning: An Overview*. <https://arxiv.org/pdf/1701.07274>.
- Li, Zewen, et al. 2020. *A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects*. <https://arxiv.org/abs/2004.02806>.
- Liu, Zhiyuan and Jie Zhou. 2020. *Introduction to Graph Neural Networks*. <https://ieeexplore.ieee.org/document/9048171>.
- Marcus, Gary. 2018. *Deep Learning: A Critical Appraisal*. <https://arxiv.org/pdf/1801.00631>.
- Mehrabi, Ninareh, et al. 2019. *A Survey on Bias and Fairness in Machine Learning*. <https://arxiv.org/pdf/1908.09635>.
- Metz, Luke, et al. 2018. *Meta-Learning Update Rules for Unsupervised Representation Learning*. <https://arxiv.org/pdf/1804.00222>.



- Mireshghallah, Fatemehsadat, et al. 2020. *Privacy in Deep Learning: A Survey*. <https://arxiv.org/pdf/2004.12254>.
- Mirsky, Yisroel and Wenke Lee. 2020. *The Creation and Detection of Deepfakes: A Survey*. <https://arxiv.org/pdf/2004.11138v1>.
- Mogadala, Aditya, et al. 2019. *Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods*. <https://arxiv.org/pdf/1907.09358>.
- Mondal, Amit. 2020. *A Survey of Reinforcement Learning Techniques: Strategies, Recent Development, and Future Directions*. <https://arxiv.org/pdf/2001.06921v1>.
- Moraffah, Raha, et al. 2020. *Causal Interpretability for Machine Learning -- Problems, Methods and Evaluation*. <https://arxiv.org/pdf/2003.03934>.
- Murshed, M.G. Sarwar, et al. 2019. *Machine Learning at the Network Edge: A Survey*. <https://arxiv.org/pdf/1908.00080>.
- Nguyen, Anh, et al. 2019. *Understanding Neural Networks via Feature Visualization: A survey*. <https://arxiv.org/pdf/1904.08939>.
- Nguyen, Thanh Thi and Vijay Janapa Reddi. 2020. *Deep Reinforcement Learning for Cyber Security*. <https://arxiv.org/pdf/1906.05799>.
- Nguyen, Thanh Thi, et al. 2020. *Deep Learning for Deepfakes Creation and Detection: A Survey*. <https://arxiv.org/pdf/1909.11573.pdf>.
- Nikolenko, Sergey. 2019. *Synthetic Data for Deep Learning Survey*. <https://arxiv.org/pdf/1909.11512.pdf>.
- O'Neill, James. 2020. *An Overview of Neural Network Compression*. <https://arxiv.org/pdf/2006.03669>.
- Oshikawa, Ray, et al. 2020. *A Survey on Natural Language Processing for Fake News Detection*. <https://arxiv.org/pdf/1811.00770>.
- Otter, Daniel, et al. 2018. *A Survey of the Usages of Deep Learning in Natural Language Processing*. <https://arxiv.org/pdf/1807.10854>.
- Pamungkas, Endang. 2018. *Emotionally-Aware Chatbots: A Survey*. <https://arxiv.org/pdf/1906.09774>.
- Peng, Huimin. 2020. *A Comprehensive Overview and Survey of Recent Advances in Meta-Learning*. <https://arxiv.org/pdf/2004.11149>.
- Pitropakis, Nikolaos, et al. 2019. *A Taxonomy and Survey of Attacks Against Machine Learning*. <https://www.manospanaousis.com/papers/pitropakis2019taxonomy.pdf>.
- Portelas, Remy, et al. 2020. *Automatic Curriculum Learning For Deep RL: A Short Survey*. <https://arxiv.org/pdf/2003.04664>.
- Puiutta, Erika and Eric Veith. 2020. *Explainable Reinforcement Learning: A Survey*. <https://arxiv.org/pdf/2005.06247>.
- Qayyum, Adnan, et al. 2020. *Secure and Robust Machine Learning for Healthcare: A Survey*. <https://arxiv.org/pdf/2001.08103>.
- Qin, Haotong, et al. 2020. *Binary Neural Networks: A Survey*. <https://arxiv.org/pdf/2004.03333>.

Rădulescu, Roxana, et al. 2020. Multi-Objective Multi-Agent Decision Making: A Utility-based Analysis and Survey. <https://arxiv.org/pdf/1909.02964>.

Raghu, Maithra and Eric Schmidt. 2020. *A Survey of Deep Learning for Scientific Discovery*. <https://arxiv.org/pdf/2003.11755>

Raisi, Zobeir, et al. 2020. *Text Detection and Recognition in the Wild: A Review*. <https://arxiv.org/pdf/2006.04305>.

Ren, Pengzhen, et al. 2020. *A Comprehensive Survey of Neural Architecture Search: Challenges and Solutions*. <https://arxiv.org/pdf/2006.02903>.

Rigaki, Maria and Sebastian Garcia. 2020. *A Survey of Privacy Attacks in Machine Learning*. <https://arxiv.org/pdf/2007.07646>.

Roh, Yuji, et al. 2018. *A Survey on Data Collection for Machine Learning: a Big Data -- AI Integration Perspective*. <https://arxiv.org/pdf/1811.03402>.

Ruder, Sebastian. 2017. *An overview of gradient descent optimization algorithms*. <https://arxiv.org/pdf/1609.04747>.

Rueden, Laura von, et al. 2019. *Informed Machine Learning -- A Taxonomy and Survey of Integrating Knowledge into Learning Systems*. <https://arxiv.org/pdf/1903.12394>.

Sagar, Ramani, et al. 2020. *Applications in Security and Evasions in Machine Learning: A Survey*. <https://www.mdpi.com/2079-9292/9/1/97/pdf>.

Sahu, Amit. 2020. *Survey of reasoning using Neural networks*. <https://arxiv.org/pdf/1702.06186>.

Sato, Ryoma. 2019. *A Survey on The Expressive Power of Graph Neural Networks*. <https://arxiv.org/pdf/2003.04078>.

Saxena, Divya and Jiannong Cao. 2020. *Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions*. <https://arxiv.org/pdf/2005.00065>.

Schuman, Catherine, et al. 2017. *A Survey of Neuromorphic Computing and Neural Networks in Hardware*. <https://arxiv.org/pdf/1705.06963>.

Sengupta, Saptarshi, et al. 2019. *Particle Swarm Optimization: A survey of historical and recent developments with hybridization perspectives*. <https://arxiv.org/pdf/1804.05319>.

Shi, Zheyuan Ryan, et al. 2020. *Artificial Intelligence for Social Good: A Survey*. <https://arxiv.org/pdf/2001.01818>.

Sloss, Andrew and Steven Gustafson. 2019. *Evolutionary Algorithms Review*. <https://arxiv.org/pdf/1906.08870>.

Song, Yangqiu and Dan Roth. 2017. *Machine Learning with World Knowledge: The Position and Survey*. <https://arxiv.org/pdf/1705.02908>.

Sorzano, C., et al. 2019. *A survey of dimensionality reduction techniques*. <https://arxiv.org/pdf/1403.2877>.

Srivastava, Yash, et al. 2019. *Visual Question Answering using Deep Learning: A Survey and Performance Analysis*. <https://arxiv.org/pdf/1909.01860>.

- Stanley, Kenneth O., et al. 2019. *Designing neural networks through neuroevolution*.  
<https://www.gwern.net/docs/rl/2019-stanley.pdf>.
- Staudemeyer, Ralf and E. Morris, 2019. *A Tutorial into Long Short-Term Memory Recurrent Neural Networks*. <https://arxiv.org/pdf/1909.09586>.
- Su, Jiawei, et al. 2019. *One pixel attack for fooling deep neural networks*.  
<https://arxiv.org/pdf/1710.08864>.
- Sun, Lichao, et al. 2018. *Adversarial Attack and Defense on Graph Data: A Survey*.  
<https://arxiv.org/pdf/1812.10528>.
- Sun, Maosong, et al. 2018. *Graph Neural Networks: A Review of Methods and Applications*.  
<https://arxiv.org/pdf/1812.08434>.
- Sun, Ruoyu. 2019. *A Survey of Optimization Methods from a Machine Learning Perspective*.  
<https://arxiv.org/pdf/1906.06821>.
- Sun, Shiliang, et al. 2019. *Optimization for deep learning: theory and algorithms*.  
<https://arxiv.org/pdf/1912.08957>.
- Thompson, Neil C., et al. 2020. *The Computational Limits of Deep Learning*.  
<https://arxiv.org/pdf/2007.05558>.
- Ucci, Daniele, et al. 2018. *Survey of Machine Learning Techniques for Malware Analysis*.  
<https://arxiv.org/pdf/1710.08189>.
- Vandenhende, Simon, et al. 2020. *Revisiting Multi-Task Learning in the Deep Learning Era*.  
<https://arxiv.org/pdf/2004.13379>.
- Vanschoren, Joaquin. 2018. *Meta-Learning: A Survey*. <https://arxiv.org/pdf/1810.03548>.
- Vinuesa, Ricardo, et al. 2019. *The role of artificial intelligence in achieving the Sustainable Development Goals*. <https://arxiv.org/pdf/1905.00501>.
- Wang, Hao, et al. 2016. *Towards Bayesian Deep Learning: A Survey*. <https://arxiv.org/pdf/1604.01662>.
- Wang, Wei, et al. 2019. *A Survey of Zero-Shot Learning: Settings, Methods, and Applications*.  
<https://dl.acm.org/doi/10.1145/3293318>.
- Wang, Yaqing, et al. 2019. *Generalizing from a Few Examples: A Survey on Few-Shot Learning*.  
<https://arxiv.org/pdf/1904.05046>.
- Wang, Zhengwei, et al. 2019. *Generative Adversarial Networks: A Survey and Taxonomy*.  
<https://arxiv.org/pdf/1906.01529>.
- Wang, Zhihao, et al. 2019. *Deep Learning for Image Super-resolution: A Survey*.  
<https://arxiv.org/pdf/1902.06068>.
- Wen, Qingsong, et al. 2020. *Time Series Data Augmentation for Deep Learning: A Survey*.  
<https://arxiv.org/pdf/2002.12478>.
- Wistuba, Martin, et al. 2019. *A Survey on Neural Architecture Search*.  
<https://arxiv.org/pdf/1905.01392>.
- Wiyatno, Rey Reza, et al. 2019. *Adversarial Examples in Modern Machine Learning: A Review*.  
<https://arxiv.org/pdf/1911.05268>.

- Wu, Zonghan, et al. 2019. *A Comprehensive Survey on Graph Neural Networks*.  
<https://arxiv.org/pdf/1901.00596>.
- Xiang, Weiming, et al. 2019. *Verification for Machine Learning, Autonomy, and Neural Networks Survey*.  
<https://arxiv.org/pdf/1810.01989>.
- Xin, Doris, et al. 2018. *How Developers Iterate on Machine Learning Workflows – A Survey of the Applied Machine Learning Literature*. <https://arxiv.org/pdf/1803.10311>.
- Yang, Shuoheng, et al. 2020. *A Survey of Deep Learning Techniques for Neural Machine Translation*.  
<https://arxiv.org/pdf/2002.07526>.
- Yao, Quanming, et al. 2020. *Adversarial Examples: Attacks and Defenses for Deep Learning*.  
<https://arxiv.org/pdf/1712.07107>.
- Yuan, Xiaoyong, et al. 2018. *Graph Neural Networks: A Review of Methods and Applications*.  
<https://arxiv.org/pdf/1812.08434>.
- Zeng, Chengchang, et al. 2020. *A Survey on Machine Reading Comprehension: Tasks, Evaluation Metrics, and Benchmark Datasets*. <https://arxiv.org/pdf/2006.11880>.
- Zhang, Daokun, et al. 2020. *Network Representation Learning: A Survey*.  
<https://arxiv.org/pdf/1801.05852.pdf>.
- Zhang, Jie, et al. 2018. *Machine Learning Testing: Survey, Landscapes and Horizons*.  
<https://arxiv.org/pdf/1906.10742v1>.
- Zhang, Quanshi and Songchun Zhu. 2018. *Visual interpretability for deep learning: a survey*.  
<https://arxiv.org/pdf/1802.00614>.
- Zhang, Wei Emma, et al. 2019. *Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey*. <https://arxiv.org/pdf/1901.06796>.
- Zhang, Yu and Qiang Yang. 2017. *A Survey on Multi-Task Learning*. <https://arxiv.org/pdf/1707.08114>.
- Zhang, Zimei, et al. 2018. *Deep Learning on Graphs: A Survey*. <https://arxiv.org/pdf/1812.04202>.
- Zhou, Jie, et al. 2019. *Graph Neural Networks: A Review of Methods and Applications*.  
<https://arxiv.org/pdf/1812.08434>.
- Zhuang, Fuzhen, et al. 2019. *A Comprehensive Survey on Transfer Learning*.  
<https://arxiv.org/pdf/1911.02685>.

**This report was written by CNA's Operational Warfighting Division**

OPS focuses on ensuring that US military forces are able to compete and win against the nation's most capable adversaries. The major functional components of OPS work include activities associated with generating and then employing the force. *Force generation* addresses how forces and commands are organized, trained, scheduled, and deployed. *Force employment* encompasses concepts for how capabilities are arrayed, protected, and sustained at the operational level in peacetime and conflict, in all domains, against different types of adversaries, and under varied geographic and environmental conditions.

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.



Dedicated to the Safety and Security of the Nation

DOP-2020-U-028073-Final

3003 Washington Boulevard, Arlington, VA 22201

[www.cna.org](http://www.cna.org) • 703-824-2000