AI Level Set

by James Guszcza and the CASBS team

Introduction

Artificial Intelligence [AI] is one of the most consequential issues of our time, but also one of the most misunderstood. Andrew Ng's slogan "AI is the new electricity" conveys the widely held view that AI is a *general-purpose technology* [GPT] – a type of technology whose breadth of applications and spillover effects can profoundly alter economies and social structures. Previous examples of GPTs include the invention of writing, the steam engine, the automobile, the mass production system, the computer, the internet – and electricity.

At the same time, there is considerable confusion about what the "AI" tagline actually means. In some contexts, it connotes the quest to build machines capable of human-level general intelligence. In others, it connotes algorithms capable only of performing specific tasks in suitably constrained environments.

The preeminent machine learning researcher Michael Jordan recently commented that "AI" is routinely used as an "intellectual wildcard" and stated,

> This is not the classical case of the public not understanding the scientists—here the scientists are often as befuddled as the public.[1]

It is therefore wise for discussions of AI governance to begin with a level-setting discussion to define terms, provide historical context, and distinguish between the various sub-concepts packed inside the "AI" tagline.

This note will take an historical approach to discuss how AI has evolved over time, and help clarify the various sub-concepts. AI experts might wish to skim or skip the expository sections of this note on first- and second-wave AI. A glossary defining common terms is provided at the end.

First-wave AI – the symbolic approach

AI is commonly agreed to date back to a conference held at Dartmouth University in the summer of 1956. The conference was convened by John McCarthy, who coined the term "artificial intelligence," characterizing it as the science of creating machines with the "ability

[1] Michael Jordan, "Artificial Intelligence – The Revolution hasn't Happened Yet." https://hdsr.mitpress.mit.edu/pub/wot7mkc1/release/8 .

to achieve goals in the world."[2] The Dartmouth Conference was attended by such AI pioneers as Claude Shannon, Alan Newell, Herbert Simon, and Marvin Minsky. Their proposal stated:

> The study is to proceed on the basis of the conjecture that *every aspect of learning or any other feature of intelligence* can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves [italics added].[3]

In other words, the original goal of AI was to implement an artificial form of human intelligence in machine form. In the essay referenced above, Michael Jordan uses the phrase *human-imitative AI* to connote this aspiration. *Artificial General Intelligence* [AGI] is also commonly used to invoke this aspiration. The AI founders were famously optimistic about their prospects of success. For example, Herbert Simon believed that human-imitative AI would be achieved by the turn of the century. Marvin Minsky wrote that, "Within a generation, the problem of creating 'artificial intelligence' will be substantially solved."[4]

These expectations reflected a philosophical belief, common at the time, that the world naturally decomposes into logical atoms, that human thought was ultimately a form of logical calculation, and that the mind is akin to a kind of "software" capable of being implemented in computers. Newell and Simon's *physical system hypothesis* ("A physical symbol system has the necessary and sufficient means of general intelligent action") reflected these beliefs.[5]

The philosopher John Haugeland dubbed this approach to artificial intelligence *Good Old-Fashioned AI* [GOFAI] and stated,

> The fundamental goal [of AI research] is not merely to mimic intelligence or produce some clever fake. Not at all. AI wants only the genuine article: machines with minds, in the full and literal sense. This is not science fiction, but real science, based on a theoretical conception as deep as it is daring: namely, we are, at root, computers ourselves.[6]

---

[2] "What is Artificial Intelligence," John McCarthy, November 2007. http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

[3] The original proposal can be found in John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon, "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence," *AI Magazine 27*, no. 4 (2006), www.aaai.org/ojs/index.php/aimagazine/article/view/1904/1802

[4] http://web.eecs.umich.edu/~kuipers/opinions/AI-progress.html

[5] Centuries earlier, Thomas Hobbes articulated a similar hypothesis in the *Leviathan*: "'Reason'... is nothing but 'reckoning,' that is adding and subtracting, of the consequences of general names agreed upon for the 'marking' and 'signifying' of our thoughts."

[6] *Artificial Intelligence: The Very Idea* by John Haugeland.

Subsequent AI developments did not live up to the early aspirations. The symbolic approach yielded such rule-based "first-wave AI" systems as the ELIZA chatbot of the 1960s and the expert systems technologies of the 1980s.[7] These systems – rigid, dependent on knowledge bases that were burdensome to create, and designed for use in specialized domains – were a far cry from the general-purpose AIs initially envisioned. As a result, the field experienced "AI winters" – periods of low enthusiasm and funding – in the 1980s and 1990s.

Second-wave AI – large-scale statistical inference

AI winter thawed into today's AI spring in the years following IBM Watson's 2011 defeat of Ken Jennings and Brad Rutter on the TV game show Jeopardy and AlphaGo's 2016 defeat of the Go grandmaster Lee Sedol. These have been heralded as watershed events in the progress of AI. Recent years have also seen our everyday personal and professional lives become increasingly suffused with AI technologies used in translation, internet search, speech and image recognition, piloting vehicles, developing new materials, drug discovery, scientific research, and statistical decision support in such domains as hiring, medical diagnosis, lending, and jurisprudence.

These developments are sometimes, at least implicitly, interpreted as evidence that we are back on the path to the human-imitative machine intelligence envisioned at the Dartmouth Conference. For example, after the AlphaGo victory, a profile of DeepMind CEO Demis Hassabis stated that:

> At DeepMind, engineers have created programs based on neural networks, modelled on the human brain. These systems make mistakes but learn and improve over time. They can be set to play other games and solve other tasks, so the intelligence is general, not specific. This AI "thinks" like humans do.[8]

---

[7] Expert systems use hand-crafted knowledge bases – sets of facts and rules – designed to assist complex decision-making in such specialized domains as law and medicine. Unfortunately, these systems proved brittle in the sense that they didn't perform well when confronted with unusual cases. Furthermore, they faced the major challenge of *knowledge acquisition*: eliciting and encoding sufficiently complete expert knowledge needed for making decisions. Expert decision-making typically involves not only on codifiable explicit knowledge, but also non-codifiable tacit knowledge. While tacit knowledge doesn't lend itself to the symbolic approach of first-wave AI, it can be imported into second-wave AI systems via the big data used to train machine learning algorithms.

[8] Demis Hassabis, master of the new machine age," Financial Times, March 11, 2016, www.ft.com/content/630bcb34-e6b9-11e5-a09b-1f8b0d268c39 . This was not an isolated statement. Two days earlier, the *New York Times* carried an opinion piece by an academic who stated that "Google's AlphaGo is demonstrating for the first time that machines can truly learn and think in a human way." Howard Yu, "AlphaGo's success shows the human advantage is eroding fast," *New York Times*, March 9, 2016. www.nytimes.com/roomfordebate/2016/03/09/does-alphago-mean-artificial-intelligence-is-the-real-deal/alphagos-success-shows-the-human-advantage-is-eroding-fast . It is also notable that DeepMind's avowed mission is to "Solve Intelligence," connoting the original goal of the Dartmouth Conference to create human-imitative machine intelligence.

Such statements obscure the true nature of today's so-called "second-wave" AI technologies. First, in contrast with the AGI aspirations expressed at Dartmouth, they are *narrow* AI technologies, capable of performing only specific tasks in suitably regularized environments. For example, an algorithm capable of recognizing human faces would be incapable of classifying tumors in medical images. An algorithm designed to drive a car would be useless at piloting a scooter – or for that matter, driving a car in a sufficiently unfamiliar environment.

More fundamentally, second-wave AI technologies essentially result from large-scale statistical inference – known as machine learning – applied to large databases. One of the most flexible and powerful machine learning techniques is known as *deep artificial neural networks*, or more simply *deep learning*. While much is made of the fact these statistical models are inspired by the networks of neurons composing the human brain, they arguably have more in common with familiar predictive algorithms – such as credit scoring models which weigh together predictive factors to guide lending decisions – than they do human brains.[9] We will briefly discuss deep learning models for illustrative purposes, not as part of an exhaustive discussion of machine learning methods. Other machine learning methods are defined in the Glossary.

An early example – the "LeNet" convolutional neural network developed at Bell Labs by Yann LeCun – was used by the US Post Office in the 1990s and 2000s to automatically recognize hand-written zip code digits. In this application, the input (predictive) variables were the pixels (either dark of light) in electronic images of hand-written digits. The outcome (target) variables – provided by humans who viewed the digits and labeled the data – were the numbers denoted by each of the images. One could imagine fitting an elementary linear regression model to this data to serve as a (very imperfect) digit classifier. The added flexibility afforded by the multi-layered structure (connoted by the word "deep") and large number of parameters in deep learning models enable such models to achieve human-level digit classification accuracy.[10]

---

[9] In an IEEE interview, the Berkeley statistician and machine learning authority Michael Jordan comments that "Each neuron [in a deep learning neural net model] is really a cartoon. It's a linear-weighted sum that's passed through a nonlinearity. Anyone in electrical engineering would recognize those kinds of nonlinear systems. Calling that a neuron is clearly, at best, a shorthand. It's really a cartoon. There is a procedure called logistic regression in statistics that dates from the 1950s, which had nothing to do with neurons but which is exactly the same little piece of architecture." Lee Gomes, "Machine-learning maestro Michael Jordan on the delusions of big data and other huge engineering efforts," *IEEE Spectrum*, October 20, 2014, http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts. For technical details relating deep learning to Generalized Linear Models (a core statistical technique which generalizes both classical and logistic regression), see Shakir Mohamed, "A statistical view of deep learning: Recursive GLMs," January 19, 2015, http://blog.shakirm.com/2015/01/a-statistical-view-of-deep-learning-i-recursive-glms

[10] The original paper LeCun et al paper can be found at: http://yann.lecun.com/exdb/publis/pdf/lecun-89e.pdf

This process of training machine learning algorithms on large bodies of human-labeled data is a form of supervised learning known as *human-in-the-loop machine learning* [HITL ML].[11] Today, thanks to the millions of electronically stored images labeled by humans, machine learning can be used to recognize images for use in facial recognition, medical image processing, piloting autonomous vehicles, and so on.[12] There is good reason for the excitement about the benefits that this technology can bring. For example, the accuracy of deep learning algorithms to classify a number of diseases using medical imaging is comparable to that of health-care professionals.[13]

It is also worth noting that HITL ML illustrates an important sense in which 2nd wave AI is parasitic on human intelligence. For example, labeling an image of a cat as a "cat," flagging a tumor "cancerous," or labeling a piece of social media content as "objectionable" all require forms of human judgment. In their book *Ghost Work*, Mary Gray and Siddharth Suri discuss the dependence of second wave AI on human labor that is often hidden. Sometimes this labor requires significant creativity or expert judgment, as in the case of flagging ambiguous medical conditions. Sometimes the labor can be harmful, as in the case of social media content moderators who suffer psychological harms.[14]

These HITL machine learning examples illustrate how statistical predictive algorithms can – unlike the rules-based first-wave AI systems – perform certain tasks (such as recognizing images or processing written or spoken language) that would otherwise require human tacit knowledge. The required tacit knowledge is encoded in the labels ("this is a stop sign"; "this tumor is cancerous.") assigned to each of the data points (a vector of pixels) used to train machine learning models. These models are then deployed as AI algorithms to make future classifications. As will be discussed shortly, human biases can also be imported into such data alongside human tacit knowledge.

Note that this discussion has focused on one form of supervised machine learning - deep learning - for the sake of illustration, rather than to provide an exhaustive discussion of the various branches of machine learning. Other major branches of machine learning include unsupervised learning, semi-supervised learning, and reinforcement learning (see Glossary). In addition, there are many forms of supervised machine learning than deep learning (examples include boosted trees and regularized regression techniques). Deep learning was selected for illustrative purposes because it is both widely used and the locus of

---

[11] Note that the more general term "human-in-the-loop" [HITL] refers to the presence of a human operator as a crucial component of an automated control process, handling such tasks as supervision and exception handling. Examples include the presence of a pilot in an airplane equipped with autopilot or the presence of a human driver in a semi-autonomous vehicle. The more specific term "human in the loop machine learning" refers to humans annotating or otherwise cleaning or adding features to the datasets used to train machine learning algorithms.

[12] In 2013, Yann LeCun's was appointed director of Facebook AI Research [FAIR]. LeCun discusses Deep Learning and Facebook's embrace of the technique in a 2015 IEEE interview: https://spectrum.ieee.org/automaton/artificial-intelligence/machine-learning/facebook-ai-director-yann-lecun-on-deep-learning

[13] https://www.thelancet.com/journals/landig/article/PIIS2589-7500(19)30123-2/fulltext

[14] For example in May 2020, thousands of moderators joined a class action lawsuit against Facebook, alleging that the job causes PTSD. https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself

considerable excitement in the AI and business communities, and because many of the conceptual and ethical issues relating to second wave AI are exemplified by deep learning.

Conceptual issues and related ethics and governance challenges

Common to rules-based first-wave AI and statistics-based second-wave AI is the creation of narrow (not general) AI systems that are capable of achieving specific goals in suitably constrained and regularized environments. The computer scientist Kristian Hammond states,

> Any program can be considered AI if it does something that we would normally think of as intelligent in humans. How the program does it is not the issue, just that is able to do it at all. That is, it is AI if it is smart, but it doesn't have to be smart like us.[15]

This definition is helpful in at least two senses. First, it establishes that applied AI should be understood in *functional* terms, rather than (as is common in the business and popular press) identified with specific machine learning or deep learning methods. This expansive, functional definition encompasses at one extreme rules-based automation systems, and at the other the deep/reinforcement learning-based technologies that dominate today's headlines.

Second, the definition makes it clear that second-wave AI should not be conflated with the original quest to implement human-like intelligence in machine form. Despite the surface similarity of neural networks to the human brain and their initially uncanny ability to perform impressive tasks ordinarily requiring human tacit knowledge, AI technologies differ from human cognition in crucial ways.

Most notably, the AI technologies that exist today or are on the horizon do not possess common sense, contextual awareness, conceptual understanding, notions of cause-and-effect or intuitive physics, theories of other minds, or the abilities to form hypotheses and reason by analogy. Unlike machine learning algorithms trained by brute force on massive datasets, human intelligence is marked by the ability to learn and abstract from very few examples. For example, even a very young child can use common sense and contextual awareness to learn a concept ("this is a chair") based on only a few instances. In contrast, deep learning algorithms do not acquire conceptual understanding and generally need to be exposed to many human-labeled examples to hopefully get it right. Similarly, a human

---

[15] https://www.computerworld.com/article/2906336/what-is-artificial-intelligence.html It might be objected that this definition is circular in the sense that it doesn't define what "intelligence" is. But as Jaron Lanier and Glen Weyl point out in their essay "AI is an Ideology, Not a Technology," the "AI" tagline "references a subjective measure of *tasks that we classify as intelligent*." For the pragmatic purposes of this discussion we find it useful to follow Hammond in using "AI" to connote the ability to achieve goals hitherto requiring what people would commonly judge "human intelligence."

driver is typically able to navigate an unfamiliar situation (e.g. a large inflatable doll blowing in the wind on a busy freeway) without the need for a plethora of prior examples.

While much is made of the near-exponential growth of computing power and data available for training second-wave AI systems, less attention is paid to a crucial form of "sparsity" inherent even in web-scale data. Big data sources typically consist of multitudinous repetitions of a relatively few common cases, such as instances of "I hope this message finds you well" in emails or images of red traffic lights captured by cameras on autonomous vehicles. But such data also tend to contain relatively few examples of innumerable rare or novel edge cases, such as a new bit of slang or an image of someone walking a bike across a multi-lane highway during unusual weather conditions. For this reason, it is often difficult to find sufficient historical examples to train machine learning algorithms to operate well in unusual circumstances.

This issue is known as "the long tail problem," connoting the unlimited variety of edge cases and unexpected scenarios (in the "tail" of the distribution of possible scenarios) that an autonomous AI system might confront. This is a major reason why brute-force machine learning is unlikely to give rise to – or replace – human-level intelligence. Even the largest datasets present limitless numbers of "small data" inference problems at which human cognition excels and machine learning algorithms choke.

In short, second-wave AI algorithms are subject to a limitation familiar from all applications of data science: They are reliable only to the extent that they have been trained on sufficient volumes of data that are suitably representative of the scenarios in which they are to be deployed (or ethically acceptable in terms of the social orders that they help create). In their book *Rebooting AI*, Gary Marcus and Ernest Davis comment:

> Without a rich cognitive model, there can be no robustness. About all you have instead is a lot of data, accompanied by a hope that new things won't be too different from those that have come before. But that hope is often misplaced, and when new things are different enough from what happened before, the system breaks down.

A neglect of this fundamental point can lead to unrealistic expectations about the capabilities of second-wave AI. For example:

- After IBM Watson's 2011 triumph on *Jeopardy*, MD Anderson Cancer Center announced a project to build an Oncology Expert Adviser using a similar approach. Perhaps under-appreciated at the time was the fact that the answers to most *Jeopardy* questions are unambiguous and electronically documented in Wikipedia pages. This contrasts with the ambiguous nature of many cancer diagnoses, and the messy and incomplete nature of US electronic health records. In 2017, MD

Anderson put the project on hold after having spent approximately $62M on it over four years.[16]

- The auto industry's overly optimistic forecasts of the arrival of fully autonomous vehicles have likely reflected a neglect of the long-tail problem and Marcus' fundamental point.[17] These problems were tragically illustrated in March 2018, when Elaine Herzbeg was killed by an autonomous test vehicle while pushing a bicycle across a four-lane road in Tempe Arizona.[18] This illustrates that data-intensive AI technologies can fail in edge scenarios that humans can handle using common sense and contextual awareness.

The second of these examples illustrates that the robustness challenges of second-wave AI can give rise to ethical issues – in this case, the possibility that autonomous systems, trained on possibly incomplete data and devoid of common sense, can cause harm. In bioethics terms, this corresponds to the principle of non-maleficence ("do no harm").[19]

More generally, the various governance and ethical issues arising from the dependence of second-wave AI on a "blank slate" approach to pattern recognition in "big data" has led some to characterize machine learning as "The High Interest Credit Card of Technical Debt"[20]: quick wins are relatively effortless and attract media coverage, while governance issues ("debt") compound silently. In addition to the robustness issue just discussed, other notable issues include:

- Algorithmic bias: Numerous examples of unacceptably biased algorithms have been documented in recent years, typically resulting from the data used to train algorithms containing patterns that correspond to societal biases. For example, a hiring algorithm built by Amazon data scientists (and never deployed) was found to be biased against female job candidates. An algorithm designed to target high-risk individuals for care management programs required blacks to be roughly twice as

---

[16] https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

[17] *Economist,* "Driverless cars are stuck in a jam," October 10, 2019; Christopher Mims, "Driverless hype collides with merciless reality," *Wall Street Journal*, September 13, 2018.

[18] https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg

[19] The core principles of bioethics are beneficence, non-maleficence, fairness, and respect for human autonomy. For discussions relating AI ethics to bioethics, see: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/6 and https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/design-principles-ethical-artificial-intelligence.html . In his above-cited essay, Michael Jordan provides an example of potential harm due to non-robust AI from the medical realm. In this case, an AI device was designed to estimate the likelihood of a fetus having Down syndrome based on ultrasound images. At a certain point, the input data's format, the resolution of the ultrasound images, changed: The AI began processing higher-resolution images to compute its estimates. This change resulted in a significant uptick in the machine's Down syndrome diagnoses. This uptick was due not to previously unrecognized cases, but to the images' higher resolution producing spurious statistical artifacts which the algorithm (trained on lower-resolution images) misinterpreted as Down syndrome indicators. It is likely that thousands of people opted for amniocentesis procedures, putting their babies at risk, based on these faulty diagnoses.

[20] https://research.google/pubs/pub43146/

sick as whites to qualify for the benefits. Numerous examples involving facial recognition algorithms have also been documented. Two examples: an internet search on "unprofessional hairstyles" yielded disproportionate images of black females; a camera face-detection program identified an Asian face as "blinking."[21]

- Susceptibility to adversarial attacks: Machine learning procedures can sometimes be tricked or "gamed" in ways that cause them to make errors or otherwise behave in undesirable ways. For example, a carefully applied bit of spray paint to a stop sign can cause a deep neural network to misclassify it as a speed limit sign.[22] Another example: in 2016, Microsoft's Tay – a chatterbot trained to interact with people based on feedback in conversations – started making racist, sexist, and authoritarian tweets within hours of being "attacked" by internet pranksters. Microsoft had to switch off the chatbot within 16 hours.[23]

- Lack of interpretability: The prominent machine learning researcher Ali Rahimi recently characterized contemporary machine learning as a form of "alchemy," meaning that researchers do not know why some algorithms work better than others, nor do they have rigorous criteria for choosing one AI architecture over another.[24] This makes it hard to characterize the conditions in which the algorithm will be reliable or error-prone. Alternately if the algorithm is intended to be used as an input into a human judgment or decision, the decision-maker might not know when or how to use a black-box indication that might not be accompanied by a plain-language "why" explanation or might come from an algorithm whose technical specifications are a trade secret. This latter issue has given rise to the area of AI research known as *Explainable AI* [XAI].

- Ethically questionable applications: Each of the above issues involve one or another form of inadequate or unsuitable data resulting in unwanted effects. AI algorithms can be intentionally or unintentionally designed in ways that cause harm or manipulate people. For example, the computer scientist Stuart Russell has warned of the possibility of a novel weapon of mass destruction: tiny weaponized drones capable of targeting people using garden variety facial recognition technologies.[25] Several examples illustrate the potential of second-wave AI to

---

[21] Amazon hiring algorithm: https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report Racially biased health benefits algorithm: https://review.chicagobooth.edu/economics/2019/article/how-racial-bias-infected-major-health-care-algorithm

[22] https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms

[23] *Tools and Weapons* by Brad Smith.

[24] https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy#

[25] https://thebulletin.org/2017/12/as-much-death-as-you-want-uc-berkeleys-stuart-russell-on-slaughterbots/ . In a commentary on Russell's "Slaughterbots" video, the defense analyst Paul Scharre characterizes it as "sensationalism" and states that while the basic concept "is grounded in technical reality," the specific nightmare scenario dramatized by Russell – drones being used as a weapon of mass destruction, killing thousands of people at a time – rests on assumptions that are "questionable, at best, to completely fanciful." Still, Scharre states that keeping the underlying technology out of the hands of would-be terrorists is a genuine problem.

manipulate people. In 2018, Google announced that its Duplex voice calling system was sufficiently lifelike to fool listeners into believing it was a human voice. After an outcry, Google clarified that the system would feature built-in disclosure. There is also concern about the potential of deepfake videos – constructed to make a person appear to do or say something that they never in fact did – to manipulate people and undermine democratic elections.[26] Another example of attempted manipulation was Cambridge Analytica's attempt to infer individuals' personality traits using social media digital exhaust in order to influence their voting behaviors.[27]

Much public discourse on the ethics and governance of AI in recent years has focused on the likelihood that AI systems might become sufficiently generally intelligent to put large numbers of people out of work, and perhaps even achieve a kind of "superintelligence" that poses an existential risk.[28] But the limitations of second-wave AI suggests that the true risks and governance challenges involve ceding too much autonomy to systems that are "intelligent" only in narrow, brittle, and sometimes harmful, ways. The computer scientist Pedro Domingos comments that, "People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world."[29]

Third-wave AI – enabling human-machine partnerships

First wave AI technologies were weak at perceiving and learning; but strong at symbolic reasoning. Second wave AI technologies are strong at perceiving and learning but weak at reasoning and explainability. Many hope that so-called *third wave* approaches to AI can build upon the strengths of both approaches, and harness insights from such fields as neuroscience, causal inference, and Bayesian probabilistic programming, to create less data-greedy systems that can better learn, reason, and promote human-machine collaboration. Barbara Grosz points out that this future trend has a venerable tradition, dating back to work in human-computer symbiosis and "Intelligence Augmentation" work of such pioneering figures as J.C.R. Licklider and Douglas Engelbart. Grosz comments that it will be important to incorporate insights from "classical AI," and comments that,

---

https://spectrum.ieee.org/automaton/robotics/military-robots/why-you-shouldnt-fear-slaughterbots . Russell, Scharre, and other coauthors subsequently collaborated on a roadmap charting a middle way between a comprehensive treaty banning lethal autonomous weapons and doing nothing for fear of foreclosing the possibility of using autonomous weapons in ways that mitigate civilian harm. https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/a-path-towards-reasonable-autonomous-weapons-regulation

[26] Google Duplex: https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update . Deepfakes: https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/

[27] https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

[28] See *Superintelligence: Paths, Dangers, Strategies* by Nick Bostrom. Bostrom states: "Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct."

[29] Pedro Domingos, *The Master Algorithm*.

> When matters of life and well-being are at stake, as they are in systems that affect health care, education, work and justice, AI/ML systems should be designed to complement people, not replace them. They will need to be smart *and* to be good teammates.[30]

Early examples of third wave AI are appearing in government funding, the business community, and academia. For example:

- The US Defence Advanced Research Projects Agency [DARPA] has announced an "AI Next" third wave AI program aimed at exploring how to give machines more human-like communication and reasoning capabilities. DARPA envisions machines that will "function more as colleagues than as tools."[31]
- Toyota Research Institute [TRI] recently announced an interdisciplinary Machine Assisted Cognition [MAC] group to explore the creation of AI tools that can better understand and predict human behavior in the context of decision-making. Perhaps not coincidentally, TRI in 2016 announced a contrarian strategy to pursue AI for "guardian angel" driver-assistance cars, rather than full "Level 5" (no steering wheel) autonomy.[32]
- Seeking inspiration from modern psychology and neuroscience, such researchers as Joshua Tennenbaum at MIT and Yejin Choi at the University of Washington attempt are exploring hybridizations of symbolic methods reminiscent of first wave AI together with deep learning models to create systems that are at once less data-hungry, more explainable, and less brittle.[33]

It is reasonable to anticipate that "smarter" and more human-compatible systems will result from expanding the paradigm of AI beyond large-scale statistical analysis to incorporate ideas and methods from other domains. At the same time, it is wise to maintain a realistic perspective of what machines likely will and will not be capable of within a practical time horizon. In his new book *The Promise of Artificial Intelligence: Reckoning and Judgment*, the University of Toronto computer scientist Brian Cantwell Smith argues that while computers will continue to outstrip human abilities at narrowly defined tasks, there is currently no scientific reason to anticipate that they will become capable anytime soon of what is he calls "judgment." Smith states,

---

[30] https://hdsr.mitpress.mit.edu/pub/wiq01ru6/release/3
[31] https://www.darpa.mil/work-with-us/ai-next-campaign
[32] TRI MAC group: https://pressroom.toyota.com/toyota-research-institute-launches-research-into-understanding-and-predicting-human-behavior-for-decision-making/ . TRI Guardian Angel car research: https://www.forbes.com/sites/roberthof/2016/04/08/toyota-guardian-angel-cars-will-beat-self-driving-cars/#6b74895e7f7f
[33] For a survey, see "AI's Next Big Leap" by Anil Anathaswamy, *Knowable Magazine,* https://knowablemagazine.org/article/technology/2020/what-is-neurosymbolic-ai?utm_campaign=newsletter-10-18-2020&utm_source=email&utm_medium=knowable-newsletter&

> Judgment requires not only registering the world but doing so in ways appropriate to circumstances. That is an incredibly high bar. It requires that a system be oriented toward the world itself, not merely the representations it takes as inputs. It must be able to distinguish appearance from reality—and defer to reality as the authority.[34]

Smith's comment strengthens the case for the third wave AI's goal of developing systems that combine the complementary strengths of human and machine intelligence.

We offer these illustrative questions to prime discussion:

1. Second-wave AI applications often rely on large bodies of data containing valuable health, economic, social, or behavioral information about large numbers of people. Many such datasets are owned by and accessible to only a small number of organizations. Given this, what ownership or wealth-sharing models can help ensure that the economic benefits of AI are equitably distributed and people's self-determination (such as the need for privacy) is not undermined?
2. Given such issues as safety and algorithmic bias, discussed above, what arrangements for safety testing, algorithmic auditing, and/or training and licencing the use of various forms of algorithms can be considered to manage the risks?
3. Given Brian Cantwell Smith's comment that while computers are adept at narrowly defined tasks but lack human judgment, what scientific principles should guide the creation of systems that optimally combine the complementary capabilities of humans and AI systems?

The Center for Advanced Study in the Behavioral Sciences is a place where great minds confront the critical issues of our time, where boundaries and assumptions are challenged, where original interdisciplinary thinking is the norm, where extraordinary collaborations become possible, and where innovative ideas are in pursuit of intellectual breakthroughs that can shape our world. CASBS @ Stanford brings together deep thinkers from diverse disciplines and communities to advance understanding of the full range of human beliefs, behaviors, interactions, and institutions. A leading incubator of human-centered knowledge, CASBS facilitates collaborations across academia, policy, industry, civil society, and government to collectively design a better future

[34] https://blogs.scientificamerican.com/observations/whats-still-lacking-in-artificial-intelligence/

Glossary

- Intelligence: The ability to solve problems, or to create products, that are valued in one or more cultural settings. (Definition from Howard Gardner – ref HCI)
- General intelligence: A person's ability to perform well on a wide range of very different cognitive tasks. Measured in psychology by Spearman's g factor. (HCI)
- Collective intelligence: Groups of individuals acting collectively in ways that seem intelligent. (HCI)
- Artificial Intelligence: The science of making computers do things that require intelligence when done by humans. Major components of artificial intelligence are learning, reasoning, problem-solving, perception, and language-understanding. (link)
- Artificial general intelligence [AGI]: Artificial intelligence with the flexibility of human general intelligence (RAI).
- Narrow AI (aka Practical AI): Artificial intelligence with the ability to achieve only specific goals (for example, piloting a car or identifying a tumor in a medical image). All past and present AI technologies far have been instances of narrow AI.
- Explicit knowledge: Knowledge that can be articulated, codified, stored, and accessed (for example in Expert Systems).
- Tacit knowledge: Knowledge that cannot be codified or expressed verbally. Characterized by Michael Polyani's slogan, "We know more than we can tell."
- Symbolic AI: Approach to AI, associated with Herbert Simon and Allan Newell, premised on the ideas that formal symbols can represent reality, and that intelligence can be reduced to symbol manipulation. Newell and Simon's physical system hypothesis states: "A physical symbol system [e.g. a digital computer] has the necessary and sufficient means for intelligent action."
- Expert system: A form of first-wave AI – computer systems designed to emulate certain aspects of the decision-making of human experts using if-then rules. The facts and rules comprising the knowledge base of an expert system are explicit knowledge.
- Robotic process automation [RPA]: The automation of mundane rules-based business processes such as insurance claims processing, sending reminder emails, screen scraping, and so on.
- Singularity: A hypothetical science-fiction scenario, first envisioned by the statistician I. J. Good, involving an "intelligence explosion" resulting from the creation of ultra-intelligent machines capable of designing even more intelligence machines.

Prominent contemporary advocates of the idea include the science fiction writer Vernor Vinge, the inventor Ray Kurzweil, and the philosopher Nick Bostrom.

- Machine learning: The development of computer algorithms that improve automatically through experience. The three major paradigms of machine learning [ML] are Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

- Supervised machine learning: A machine learning paradigm in which a function is optimized to map an input (e.g. a collection of vectors of 0s and 1s, each denoting a pattern of black or white pixels in a corresponding collection of photographs) to an output (e.g. a corresponding collection of labels indicating whether the photographs are of cats or dogs). When the labels are supplied by human workers, the process is known as human-in-the-loop machine learning. The collection of inputs and outputs is known as training data, used to mathematically estimate the parameters of the function. Colloquially such machine learning-derived functions, when implemented as pieces of AI software, are typically referred to as "machine learning algorithms" or "AI algorithms."

- Unsupervised machine learning: A machine learning paradigm containing numerous techniques to detect interesting or meaningful patterns in a dataset without the use of the output labels characteristic of supervised machine learning. Examples include clustering methods, market basket analysis ("customers who purchase x tend also to purchase y"), and such dimension reduction methods as principal components analysis.

- Reinforcement Learning: A machine learning paradigm in which software agents are trained by trial and error to make sequences of decisions that maximize a notion of cumulative reward, using a feedback system of punishments and rewards. Such agents must trade off exploring uncharted territory with exploiting current knowledge.

- Artificial neural network: A form of machine learning loosely inspired by the biological neural networks in animal brains.

- Deep learning: A type of machine learning, based on artificial neural networks, that enable computer systems to automatically discover representations needed for feature detection or classifications (e.g. "whisker" or "cat") from raw data (e.g. pixels in photographs). The term "deep" connotes not psychological "depth" but the presence of multiple "hidden layers" in the neural network architecture.

- Generative Adversarial Network [GAN]: A system of two artificial neural networks designed to contest each other in a zero-sum game (in which one agent's gain is the other's loss). One of the networks, called the "generator," generates new raw data with similar statistics as the training dataset in an attempt to fool an evaluating network, called the "discriminator." The GANs trained on photographs can generate new photographs that can look authentic to human observers.

- Deepfake: Synthetic photographs or videos created using Generative Adversarial Networks [GANs], possibly for such sinister purposes as falsely incriminating people or creating fake social media profiles.

- Deep Reinforcement Learning: A technique developed at DeepMind that combines deep learning for recognizing patterns with reinforcement learning for learning based on reward feedback signals. Deep reinforcement learning famously enabled AlphaGo to meet the world Go champion Lee Sedol by generating millions of self-played games of go. Previously the space of combinatorial possibilities in Go had been considered too large for machine learnings to effectively learn in. Deep reinforcement learning is well suited to learning in high-dimensional environments described by a fixed set of rules, such as strategy or video games.

- Natural Language Generation: A field at the intersection of linguistics, computer science, and AI concerned with analyzing and processing large amounts of natural language data. Subfields include speech recognition, natural language understanding, and natural language generation. GPT-3, developed by OpenAI, is a deep learning-based natural language generator capable of producing human-like text.

- First-wave AI: See symbolic AI

- Second-wave AI: The development of AI technologies using large-scale statistical inference, in particular forms of machine learning such as Deep Learning and Reinforcement Learning.

- Third-wave AI: DARPA characterizes third-wave AI as the quest to "[transform] computers from specialized tools to partners in problem-solving." Aspects of this include creating AI that is more transparent or explainable, capable of learning from few examples, and helping humans overcome natural cognitive limitations.