AUGUST 2021

The Spectrum of Artificial Intelligence

Companion to the FPF AI Infographic

FUTURE OF PRIVACY FORUM

AUTHORED BY

Brenda Leong
Senior Counsel & Director of Artificial Intelligence and Ethics

Dr. Sara R. Jordan
Senior Researcher, Artificial Intelligence and Ethics

COVER ART

Artist: Dr. Lydia Kostopoulos

Title: Luncheon of the Tech Enabled Boating Party

Year: 2020

Contents

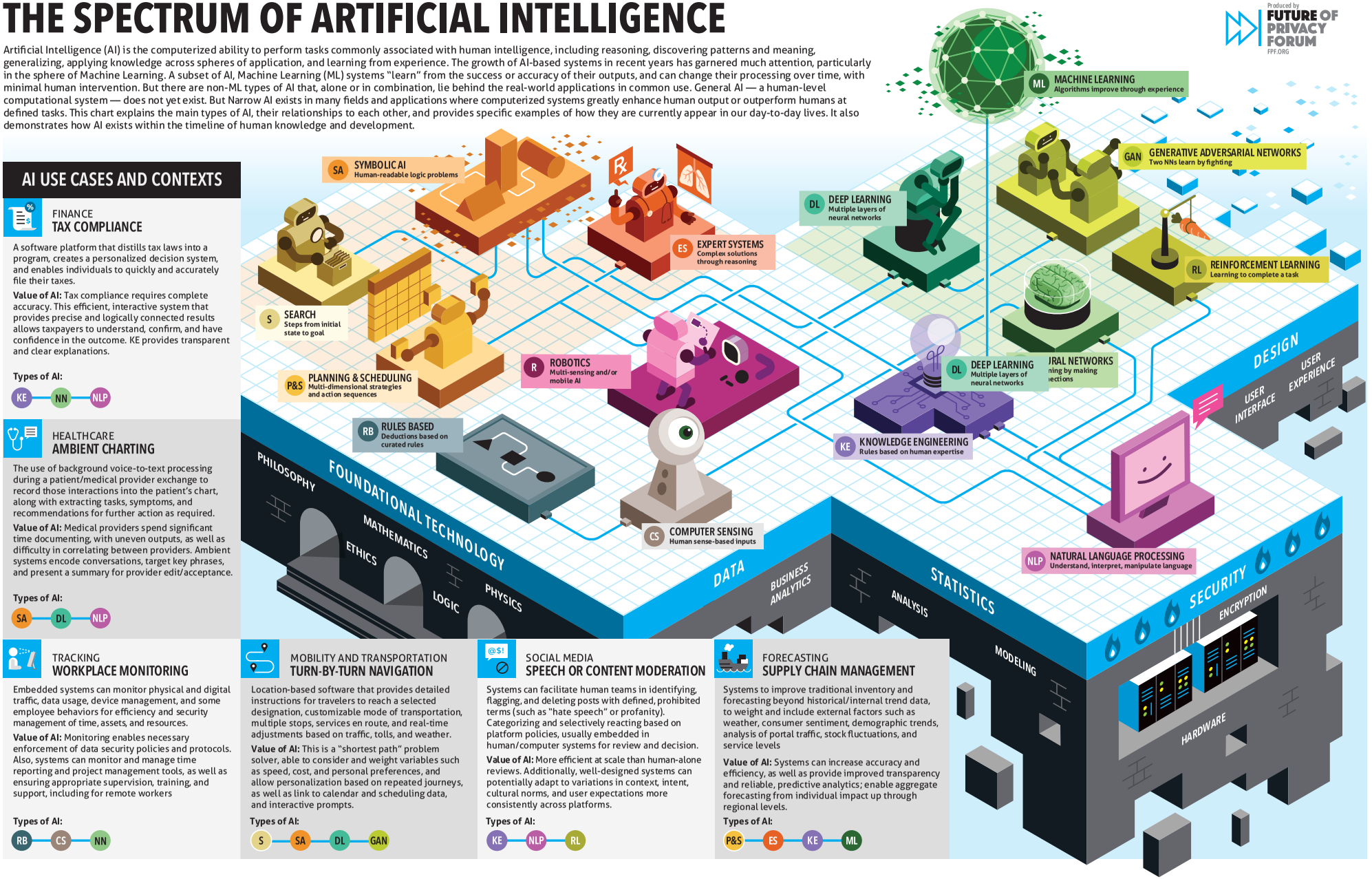
The Spectrum of Artificial Intelligence

Companion to the FPF AI Infographic

THE SPECTRUM OF ARTIFICIAL INTELLIGENCE

Produced by FUTURE OF PRIVACY FORUM FPF.ORG

Artificial Intelligence (AI) is the computerized ability to perform tasks commonly associated with human intelligence, including reasoning, discovering patterns and meaning, generalizing, applying knowledge across spheres of application, and learning from experience. The growth of AI-based systems in recent years has garnered much attention, particularly in the sphere of Machine Learning. A subset of AI, Machine Learning (ML) systems "learn" from the success or accuracy of their outputs, and can change their processing over time, with minimal human intervention. But there are non-ML types of AI that, alone or in combination, lie behind the real-world applications in common use. General AI — a human-level computational system — does not yet exist. But Narrow AI exists in many fields and applications where computerized systems greatly enhance human output or outperform humans at defined tasks. This chart explains the main types of AI, their relationships to each other, and provides specific examples of how they are currently appear in our day-to-day lives. It also demonstrates how AI exists within the timeline of human knowledge and development.



AI USE CASES AND CONTEXTS

FINANCE
TAX COMPLIANCE

A software platform that distills tax laws into a program, creates a personalized decision system, and enables individuals to quickly and accurately file their taxes.

Value of AI: Tax compliance requires complete accuracy. This efficient, interactive system that provides precise and logically connected results allows taxpayers to understand, confirm, and have confidence in the outcome. KE provides transparent and clear explanations.

Types of AI: KE — NN — NLP

HEALTHCARE
AMBIENT CHARTING

The use of background voice-to-text processing during a patient/medical provider exchange to record those interactions into the patient's chart, along with extracting tasks, symptoms, and recommendations for further action as required.

Value of AI: Medical providers spend significant time documenting, with uneven outputs, as well as difficulty in correlating between providers. Ambient systems encode conversations, target key phrases, and present a summary for provider edit/acceptance.

Types of AI: SA — DL — NLP

TRACKING
WORKPLACE MONITORING

Embedded systems can monitor physical and digital traffic, data usage, device management, and some employee behaviors for efficiency and security management of time, assets, and resources.

Value of AI: Monitoring enables necessary enforcement of data security policies and protocols. Also, systems can monitor and manage time reporting and project management tools, as well as ensuring appropriate supervision, training, and support, including for remote workers

Types of AI: RB — CS — NN

MOBILITY AND TRANSPORTATION
TURN-BY-TURN NAVIGATION

Location-based software that provides detailed instructions for travelers to reach a selected designation, customizable mode of transportation, multiple stops, services en route, and real-time adjustments based on traffic, tolls, and weather.

Value of AI: This is a "shortest path" problem solver, able to consider and weight variables such as speed, cost, and personal preferences, and allow personalization based on repeated journeys, as well as link to calendar and scheduling data, and interactive prompts.

Types of AI: S — SA — DL — GAN

SOCIAL MEDIA
SPEECH OR CONTENT MODERATION

Systems can facilitate human teams in identifying, flagging, and deleting posts with defined, prohibited terms (such as "hate speech" or profanity). Categorizing and selectively reacting based on platform policies, usually embedded in human/computer systems for review and decision.

Value of AI: More efficient at scale than human-alone reviews. Additionally, well-designed systems can potentially adapt to variations in context, intent, cultural norms, and user expectations more consistently across platforms.

Types of AI: KE — NLP — RL

FORECASTING
SUPPLY CHAIN MANAGEMENT

Systems to improve traditional inventory and forecasting beyond historical/internal trend data, to weight and include external factors such as weather, consumer sentiment, demographic trends, analysis of portal traffic, stock fluctuations, and service levels

Value of AI: Systems can increase accuracy and efficiency, as well as provide improved transparency and reliable, predictive analytics; enable aggregate forecasting from individual impact up through regional levels.

Types of AI: P&S — ES — KE — ML

I. Executive Summary

This paper is a companion piece to the FPF Spectrum of Artificial Intelligence (AI) Infographic, to expand the information included in that educational resource, and describe how the graphic can be used as an aide in developing legislation or other regulatory guidance impacting AI-based systems. We identify specific use cases for various AI technologies and show how the differing algorithmic architecture and data demands present varying risks and benefits. We discuss the spectrum of algorithmic technology and demonstrate how design factors, data use, and model training processes should be considered for specific regulatory approaches.

Recent calls for regulation of AI-based systems exist within the complex landscape of contemporary AI programs within a variety of industries and applications from agricultural crop growth detectors to automated book recommendations.[1] Some approaches focus on a need to cover "algorithms" generally while other initiatives are narrower, suggesting regulation of AI specifically as used in automated vehicles[2] or medical devices[3]. This paper aims to assist the development of any of these approaches by making clear some of the terms, relationships, and functions of AI, and how they might be impacted by differing restrictions. Any regulatory efforts should be made with the best understanding of how different systems, use cases, applications, and ultimately individuals, will be affected. Despite the media focus on Machine Learning (ML), there are many other types of AI preceding and operating alongside ML, all of which have different attributes and aspects, and which will need differently formulated regulatory controls or guidance.

Many forms of artificial intelligence are the clear analogue to human thought processes or human physical actions which can be integrated within systems to help humans process or move faster, more precisely or consistently, and with more information than an individual could. But, of course, AI systems can also process large amounts of data beyond what any human could do, to identify patterns, make connections, and predict outcomes that are far beyond what conventional data science and statistical methods could accomplish.

A. Symbolic AI

Traditional Rules-Based AI leading to Symbolic AI (sometimes used synonymously) is the collection of all methods in artificial intelligence research that are based on human-readable representations ("symbols") of mathematics, logic, and coded programming. Symbolic AI systems represent the first significant steps towards designing machines to enable complex decisions or reason through complexity and uncertainty. There are several algorithmic designs that are generally considered to be examples of Symbolic AI, including Search, Planning and Scheduling, and Expert Systems.

When computers are programmed to find a specific pattern in a set of symbols and then to perform a designated action, we can say that the AI is engaged in a "**search**." **Planning and scheduling** AI are what enable a computer to take into account multiple dimensions that require adjustments of strategies, such as collecting tokens in a video game (e.g. gaining points) while avoiding traps (e.g. losing lives). Search AI and planning and scheduling AI play important roles in the boring and hidden parts of systems that provide many of the conveniences in modern life. For example, supply chain management, including airline cargo scheduling systems and "just in time" restocking models rely on these programs. **Expert systems** identify solutions by combing through and combining multiple types and layers of information, using the reasoning and logic common to a particular profession or specialty, such as medicine or engineering. An expert system can provide faster, more reliably accurate diagnoses based on personal health data, or design recommendations for pollution abatement given the environmental and industrial factors involved.

Building computers that can "see," "listen," "smell," or "taste" as ways to evaluate their physical environment requires new approaches to **computer sensing** that generally rely on a combination including several other forms of AI as well. Computer sensing plays an important role as a building block for creating augmented intelligence used in advanced and assistive **robotics.**

Teaching AI to reason using the same cognitive patterns as expert professionals involves teaching machines the bases of professional common knowledge laid out as a set of underlying rules for processing and establishing expectations, knitting together the wealth of documents, rules, and common knowledge of professions. Known as "**knowledge engineering**" systems, they use rules and pattern recognition to sift through data such as tax codes, extract relevant patterns, and categories, and provide an expert's guidance. For example, this analysis could formulate question and answer sets to guide an individual through their tax return preparation, following the appropriate steps for that individual's finances.

Natural language processing (NLP) systems are some of the most common AI systems that people routinely encounter. Powering home-based assistants, and various devices or appliances; providing language translation tools, predictive typing, autocorrect, question and answer systems, and robotic speech; summarizing and analyzing written texts, and comparing drafts for plagiarism; and even writing independent, creative work, NLP combines ML with other forms of AI in everyday systems and tools.

B. Machine Learning

The types of AI described above tell computers how to sift through information according to rules and processes crafted by humans, such as language or mathematics. **Machine learning** is different. Machine learning works because machines use an initial set of rules (programming) to identify connections and patterns which they then use to internally edit their instructions or build additional rules of their own. Computers using the results of prior analyses to improve subsequent calculations or minimize loss of performance is what makes machine learning so powerful. Machine learning has improved traditionally difficult AI tasks, such as image recognition, and provides the ability to analyze constantly changing information flows for applications like social media content monitoring.

Neural networks, the building blocks of some types of machine learning, learn by identifying patterns within input data to make new, internal, rules about the relationships between the data and outputs. These systems allow computers to process highly complex information quickly, sometimes approaching human levels of association and "intuition." These networks can be layered to process data through multiple programs sequentially and repeatedly for more sophisticated analysis. When they are layered, they may comprise a "**deep learning**" system. These are the systems trained to recognize objects in photos, evaluating for color values, edges, and commonly associated items so that the output value, such as probability that a specific image is a canoe and not a cat, can be provided to the user.

Most recently, there are two newer forms of machine learning enabling powerful systems to achieve major advances: **generative adversarial networks (GANs) and reinforcement learning (RL)**. Reinforcement learning is a key step in designing AI to independently learn human-like, goal-oriented, tasks. Reinforcement learning is what powered the system that learned to master the game "Go." The first ML system, which defeated the human world Go champion 4 games to 1, operated on traditional machine learning processes. The next program was designed using reinforcement learning and beat that original system, 100-0. Reinforcement learning systems will likely power the next generation of robotics, for purposes such as search and rescue missions in complex environmental situations or high-capability home care assistants.

GANs, the newest variation of machine learning AI being developed, are based on creating a pair of neural networks that learn by attempting to better each other: first, the "generator" of the pair creates an output (e.g., an image) based upon the initial human programming. The other network, the "discriminator," has been programmed to what the correct output should be (e.g., what the image should look like). The discriminator evaluates the output, and critiques it. Initial outputs are likely to be extremely inaccurate. The discriminator's feedback is then incorporated, the generator continues to churn out results, and the feedback loop continues until the generator produces data that the discriminator believes meets the quality expectations. This GANs type of learning is what drives "deep fakes" and some entertainment uses of AI and will likely inform or improve other systems in the future.

C. Risks and Benefits of AI

The future of AI ultimately lies in the goals and systems towards which humans direct it. Responsible uses should include two primary foundations for AI: to further advance human knowledge and to improve human lives. AI is key to the future of knowledge in many scientific disciplines and commercial technologies but carries accompanying risks that it will be applied unethically, or designed unfairly, and that individuals and groups will be worse off in specific or personal ways. However, the potential benefits are powerfully significant, if sufficient effort is applied to ensure fair and beneficial impacts for a greater social good.

AI systems operate across a broad spectrum of scale. Processes using these technologies can be designed to seek solutions to macro level problems like environmental challenges around undetected earthquakes, pollution control, and other natural disaster responses while they are also incorporated into personal level systems for greater access to educational, economic, and professional opportunities. If regulation is to be effective, it should focus on both technical details and the underlying values and rights that must be protected from adverse uses of AI, to ensure that AI is ultimately used to promote human dignity and welfare.

II. Introduction

This paper is a companion piece to the FPF Spectrum of Artificial Intelligence (AI) Infographic, to further explain the information included in this educational resource, and in particular how it should be used as an aide in developing any legislation around AI-based systems.

If one of the many calls to regulate AI were successful, what exactly would be regulated? In many ways, the answer to this question is "it depends." Some approaches start from the premise that AI is a discrete type of software technology, like an operating system, and assume regulations should focus on preventing unfair impacts on things like hiring or credit scoring. Other approaches define AI as an entire system of hardware plus software, like a robotic arm or a personal home assistant, and should be regulated in ways based on safety concerns, similar to modern connected automobiles. As we will discuss, the contemporary spectrum of AI is broad, and any call to regulate AI must align regulatory controls that are appropriate to the context, and in light of the specific harms of the systems being considered.

Artificial intelligence is a term with a long history. Meant to denote those systems which accomplish tasks otherwise understood to require human intelligence, AI is directly connected to the development of computer science but is based on a myriad of academic fields and disciplines, including philosophy, social science, physics, mathematics, logic, statistics, and ethics. AI as it is designed and used today is made possible by the recent advent of unprecedentedly large datasets, increased computational power, advances in data science, machine learning, and statistical modeling. AI models include programming and system design based on a number of sub-categories, such as robotics, expert systems, scheduling and planning systems, natural language processing, neural networks, computer sensing, and machine learning. In many cases of consumer facing AI, multiple forms of AI are used together to accomplish the overall performance goal specified for the system. In addition to considerations of algorithmic design, data flows, and programming languages, AI systems are most robust for use in equitable and stable consumer uses when human designers also consider limitations of machine hardware, cybersecurity, and user-interface design.

Two arguments are most commonly behind calls for regulation. The first is that AI presents unique risks to humans, the environment, or social and cultural values at a scale and scope beyond prior technological advancements. Under this argument, AI regulations should introduce mechanisms to identify and mitigate potentially harmful outcomes, both tangible and intangible, as well as offer solutions to rectify situations where poorly managed or unforeseen risks caused damage to humans, the environment, or social good.

The second, narrower argument focuses on the unknown risks that may arise due to the innate design aspects of AI that appear to be out of the control of the programmer or user of the system. Under this argument, because some forms of AI are designed in a way that humans cannot fully comprehend, explain, or reproduce, we should use AI with regulatory-driven caution until we can be more confident of the reliability or accuracy of these systems. Both of these arguments assume there are unique aspects to AI-based systems that are not, or cannot, be addressed by existing legal protections, regulatory schemes, or traditional policy approaches.

Our purpose is not to assert if or how AI systems should be regulated, but rather to provide a general understanding of the variety of AI systems that may be behind various applications and industries, and what the impacts might be, and to demonstrate that any regulatory action should be taken thoughtfully and in response to the particular areas of concern. Blunt approaches that seek to include any and all systems using AI-based models are considerably more likely to have undesired impacts, and/or unforeseen secondary consequences.

Toward that end, this paper outlines the spectrum of AI technology, from rules-based and symbolic AI to advanced, developing forms of neural networks, and seeks to put them in the context of other sciences and disciplines, as well as emphasize the importance of security, user interface, and other design factors. Additionally, we seek to make this understandable through providing specific use cases for the various types of AI and by showing how the different architecture and data demands present specific risks and benefits.

Across the spectrum, AI is a combination of various types of reasoning. Rules-based or Symbolic AI is the form of algorithmic design wherein humans draft a complete program of logical rules for a computer to follow. Newer AI advances, particularly in machine learning systems based on neural networks, are able to power computers that carry out the programmer's initial design but then adapt based on what the system can glean from patterns in the data. These systems can score the accuracy of their results and then connect those outcomes back into the code in order to improve the success of succeeding iterations of the program.

A commonly used comparison for an algorithm is that it is like a recipe. This analogy applies well to rules-based and symbolic AI. Like a recipe instructs the cook to combine specific ingredients in a specific order to make a

particular dish, a rules-based AI system is a set of commands such as if-then logic statements that tell a computer how to combine precise forms of information to achieve a particular task or outcome. Algorithms can be more or less exact in their outputs based upon the clarity of the instructions we give. Just as a recipe written by a chef may call for a "pinch of salt" because most chefs have a common understanding of what a "pinch" is, an AI system that instructs a computer to reason probabilistically or approximately may be used in cases where a "good enough" answer helps augment human decision-making. Also like recipes, AI systems can range from simple to extremely complex, with some of the most advanced systems being combinations of multiple algorithms creatively combined.

III. Foundation Disciplines[4]

AI and Machine Learning (ML) are relatively new features of the scientific landscape, but they are built upon a long history of philosophical and scientific developments, including areas such as philosophy, ethics, logic, mathematics, and physics. In more recent decades, scientific disciplines that inform AI include data analytics, statistical modeling, and cybersecurity and encryption. Furthermore, these systems cannot be evaluated without also including considerations about the hardware devices and networks, the underlying logics and principles, and user interface and user experience designs. All of these areas contribute to the questions, answers, and analysis needed to fully review present day AI. While some might seem remote, or tangential, in fact the underlying values and assumptions of the designers and users are key to understanding the contextual implications of any particular AI system.

Philosophy

Each of the foundational disciplines of artificial intelligence discussed on the following page — logic, mathematics, physics — were once part of what we now call philosophy. In Western tradition, philosophy is the love of wisdom; per the Greeks, to love wisdom is to seek the pursuit of knowledge. While knowledge can be both theoretical or practical, within the context of modern AI, the impulse to gain practical knowledge to deploy to a particular goal generally dominates.

Western philosophy is only one of the views that informs thinking about "intelligence" in computer systems. The Indian subcontinent describes philosophy as a debate between humans, nature, and the gods. Wisdom through their texts is the ability to deploy wit alongside theoretical or eternal truths on the way to gaining a full grasp of being, knowledge, and even nothingness. The competing traditions of Chinese philosophies reflect the salience of learning, character development, an appreciation of uncertainty, and the importance of meaning as the product of a lifetime of seeking. Other philosophical traditions also bear upon how we measure the intelligence of AI systems today.[5]

Ethics

Ethics is a potentially controversial term with both positive and negative connotations. Historically, ethics is a discipline that trains minds to consider universal ideas about what it means to be human, and addressing questions such as "what is good, how might I be good, or how might I do good?" But ethics can also generate concerns about people shifting boundaries around good and evil, or subjectively defining right or wrong toward a particular agenda or outcome.[6] The vagueness about what is or isn't included in ethics has also given rise to the use of synonyms and associated concepts for this area, such as "social responsibility," "morals," "values," or "human rights" without always being clear or consistent as to what each of these means. These various perspectives on ethical considerations have flowed over into twenty-first century discussions around the appropriate roles of industry, technology, and automation, all of which continue to influence contemporary debates about ethics in AI. These now include robust arguments of ideas such as whether an AI system can "be" ethical (or unethical), whether it's dependent on the applications and context, whether human labor and safety priorities should pre-empt technological efficiencies, and even ideas like whether robots are independent entities with agency or rights.[7]

IV. Modern Components

A. Data

"Data" is the organized record and presentation of information. Modern "big data"[12] in the digital world is part of the trail created by people, businesses or other organizational entities as they interact with sensors, surveys, sites, and applications embedded in the array of internet or computer-based, or computer-connected, products, services, and features.[13] Data in this context has measurable characteristics like volume, variability, accuracy, and value, and also can have organizing principles around protection, privacy, provenance, and proportionality.[14] In discussions about AI and ML, personal data is often categorized according to the sector from which it is gathered, such as healthcare, financial, educational, or consumer data. Data is also described based on the technological associations, including sensor data, location data, or visualized data.[15] Data can also be synthetic, imputed, or meta (data about data). Data is not just the invaluable input into many AI and ML systems, but also includes the information created by or describing the analysis inherent in those systems, and the outputs and recommendations the systems produce.

The techniques necessary for processing data for uses in AI and ML have spurred both theoretical and applied research in areas of data management, including the ethical processing of data (e.g., differential privacy) and the technical processing techniques like dimensionality reduction (e.g., principal component analysis).[16] The demand for AI-ready data has also spawned an industry for data engineering, a new form of AI-adjacent specialization that serves to expedite, but potentially complicate, explainability and transparency for new machine learning applications.

Cleaning data, identifying bias in data, adding noise to data, deidentifying data, standardizing machine-readable data, analyzing data, and "owning" data are just a few of the many issues surrounding the collection, flow, and use of all of this information in modern digital systems.

Logic

Logic is a branch of mathematics based on established rules and proofs. It is concerned with evaluating conditions and relationships with statements such as "if, then." Symbolic logic, modal logic, predicate logic, syllogistic logic, and computational logic are all part of the study of how humans record their ideas, reason about their innate connections, and reach conclusions that can be replicated.[8] In the areas of AI and machine learning, logic plays an intrinsic role in the translation of human rules, conclusions, or understanding into machine parameters and actions. Indeed, symbolic and rules-based AI could not exist in their known forms without logic.[9]

Mathematics

Outside of computer programming, AI and machine learning can almost easily be defined as areas of applied mathematics.[10] Weaving together linear algebra, calculus, combinatorics, graph theory, information theory, probability, and vector analysis, machine learning in particular depends on the mathematical manipulation and analysis of data. Mathematical thinking and the representation of the world in equations and numerical terms are critical to explanations of AI and machine learning.

Physics

Even without consideration of the emerging, transformational impact on AI that may arise when quantum computing becomes more widely understood and available, the role of physics in artificial intelligence cannot be understated.[11] Problems associated with mining the data from massive physics research projects (e.g., CERN, astronomy) have pressed machine learning experts to identify new methods for managing data at the scale of the universe, as well as that of the quant. At the end of the day, AI systems operate in the physical world and are bound and scoped by the physical restraints of human-scaled applications.

B. Statistics

A branch of applied mathematics often tailored into specific fields (e.g., statistical mechanics, biostatistics), statistics inform the way in which the inputs for AI systems are evaluated and how the outputs are interpreted.[17] Statistical modeling is the design of a way to make sense of data through analysis. Without established statistical theory, we would lack many of the essential concepts to explain how AI works or to evaluate how accurately machines have identified patterns. Established tools and measures such as variance, correlation, confidence, probability, and hypothesis testing inform the algorithmic models in order to glean insights and inform evidence-based reasoning.[18] This reasoning drives how systems combine data to test representations of reality against one another, and provides the tools necessary to evaluate the results.

C. Design

While not always the first area of consideration for AI systems, the physical design of AI systems includes considerations of aesthetic value, creativity in physical presentation, and the efficiency of human interface. Apple, for example, is particularly known for its focus on beauty and intuitive function as part of its innate design structure.[20] "Design thinking" — a process for brainstorming collaboratively — includes 5 stages: Empathize, Define, Ideate, Prototype and Test and is deployed widely in the software and applications development landscape in user interface design and user experience design. More than simple efficiency or functionality, however, design considerations for AI systems must also include the understanding that they are integral components of applications that affect the lives of individual people, communities, cultures, and the environment. Whether focusing on personal privacy, bias and fairness, civil rights impacts, or other social impacts, design choices are a fundamental consideration in any AI system.

Security

Security is an intrinsic element in the trustworthiness and reliability around AI systems, including hardware, software, and network components. AI can be both a vulnerability, and a part of the defense of computers systems against cyber intrusion. Data can be "poisoned" at multiple stages in the life of a model, and models themselves can be subverted. But AI-based systems are also being integrated as powerful tools for defending against cyber attacks targeting datasets as well as existing systems, such as energy grids, or defense networks. Security concerns are a critical component of all stages of system design and implementation. In addition to the traditional threats to data, networks, and storage, AI models in particular are at risk from inference, inversion, and extraction attacks — all of which potentially compromise the system, whether by revealing the data, or inappropriately influencing the outcomes or future processing.

Hardware

From the physical building of "server farms," to considerations of the environmental impact of the their huge power requirements, AI systems occupy a substantial place in the physical world, integral to providing the digital and virtual platforms we have all come to take for granted.

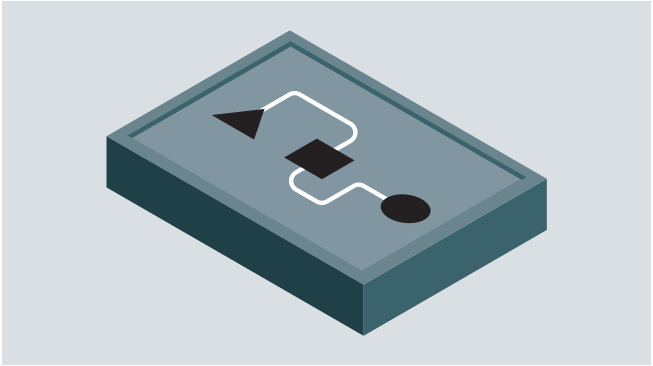
Modern AI, and in particular ML, are possible only because of advances in computer processing speeds and associated hardware infrastructure, such as high-density, environmentally controlled servers, high-resolution monitors, and even peripherals like HDMI cables. Progress in AI depends on continuous, reliable access to high performance computing hardware and transmission channels. Even for cloud applications or virtual environments, there is a hardware infrastructure where applications reside, computations are performed, and system management carries out the necessary planning in support of efficient distribution of AI processing.[19]

V. Artificial Intelligence – Overview

Artificial intelligence covers many combinations of hardware and software components, but at its core, means the set of actions that would normally be understood to require human intelligence. The most dazzling examples of AI that appear in science fiction are identical or superior to generalized human cognitive and intuitive powers. Whether this type of AI, known as General (or Strong) AI is even possible remains speculative, with no clear idea of when such capability might develop. However, Narrow (or Weak) AI is defined as the operation by systems performing particular functions in a specific context, application, use case, or circumstance. Thus, the greatest chess player, and the world Go champion, are now AI systems. These systems, while far outstripping human capabilities in one specific area, are fairly useless for almost any other purpose. While some aspects of their programming and sub-routines might be reused to jump start other projects, it is not a system that can intrinsically "learn" a new skill (function or application) without human involvement to edit and reapply that code, provide new data and training, and so on.

Across the spectrum, AI is a combination of various types of reasoning. Rules-based or Symbolic AI is that form of algorithmic design wherein humans draft a complete program of logical, connected commands for a computer to follow.[21] Newer AI advances, particularly in machine learning systems based on neural networks, are able to power computers that carry out the programmer's initial commands but then adapt their operations based on what the system can glean from patterns in the data. These systems evaluate their results and then connect those outcomes back into the code in order to improve the success of succeeding iterations of the program.[22]

Recently, what the media and many discussions around AI are most commonly referring to is actually machine learning, as if those two terms were exactly interchangeable. In fact, however machine learning as a specific subset of AI. Explaining the specifics of machine learning and the distinction between it and other types of AI is one of the primary goals of this paper and infographic. All ML is a form of AI, but not all AI is ML. This is one of the key takeaways to inform potential policymakers in their consideration of regulating AI systems — that they should correctly identify what type of systems are behind the functions and applications they are concerned with, and how those operate, so that standards, guidance, and restrictions can be targeted appropriately.



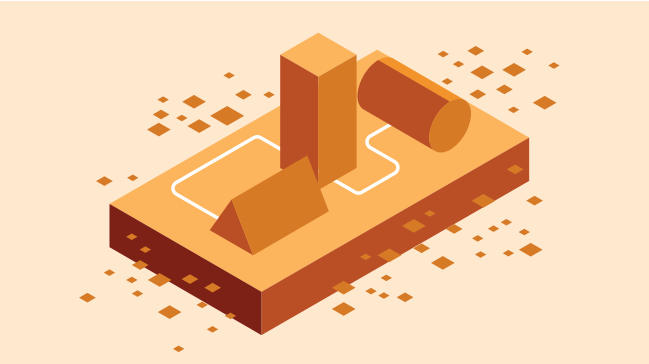
A. Rules-Based AI

Artificial intelligence has been part of the programming landscape, built on the academic foundational disciplines described above, since at least the 1940s. The earliest forms of AI are still in use today, including within some of our most widely used systems. In fact, examples of rule-based AI, such as those which make up the backbone of navigation systems, are so commonplace we often do not think of them as artificial intelligence any longer.

The initial drive to build artificial intelligence grew from the human desire to automate work that is repetitive, tedious, dangerous, requires high levels of precision, or is simply impossible for an individual or group of humans given some combination of factors. Many of these tasks are dependent on *heuristics* — formal and informal rules — that "everyone" within that field or domain just "knows." Some of the first examples of rules-based or heuristic AI emerged in fields like chemical analysis, infectious disease diagnostics, or oncology diagnostics. What was relevant about these domains was the belief that the logical structure of the questions to be asked and the specific types of information necessary to answer those questions was both available and could be symbolically represented to a computer. The rules for reaching "good" decisions, and the expert knowledge of uncertainty or confidence in the utility of a specific piece of information, could be systematically organized and ranked to produce a reliable, consistent answer by a non-human program that was on par with, or superior to, that deduced or provided by the human expert.

Rules-based systems maneuver through data using an "inference engine";[23] a set of logical commands according to which a computer interprets information and relates that information to the set of possible outputs (e.g., a probability score) in order to answer the questions asked of it. Because of the logic and symbolic structure of inference engines driving rule-based systems, they can work both backwards and forwards.

Inference engines work by either "forward chaining" or "backward chaining" through the rules. In a forward chaining strategy, an inductive approach (moving from individual facts to general theory) moves forward from data inputs to conclusory outputs by matching the data to the best rule to identify a match. In a backward chaining strategy, a deductive logical approach (from general theory to individual facts) moves backward from the provided set of possible outputs, through the rules and data, to identify if a particular output can be supported as legitimate. Where a known set of possible outputs or decisions is generally well characterized, such as in healthcare diagnostic systems, a backward chaining strategy is useful. Where there is a wider or uncertain range of possible outputs from a system such as selecting, combining, and providing the appropriate mix of medications, a forward chaining strategy is useful.



B. Symbolic AI

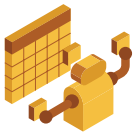
The terms Symbolic AI and Rules-based AI are many times used interchangeably. Systems that automate processes in an imitation of human reasoning, such as through the use of logical statements or identification of patterns, are a type of symbolic AI. Computers can be designed to detect patterns such as sequences of letters, numbers or other symbols, and then to reproduce an arrangement of those symbols so that humans can evaluate them. There are various versions of this type of system, which we are using as the umbrella term for the following three sub-categories:

i. Search

Search algorithms are used by many people in contemporary life. Since the beginning of the age of internet search engines, "search" has come to be synonymous with online browser search capabilities, but that is not the entirety of what this category includes. Search algorithms are used in many systems that are designed to find the best path to a goal within a specific set of possible solutions. Search algorithms help us to achieve goals such as reaching the highest score on a game, finding the shortest or fastest route to a destination, or finding a recipe for vegan chocolate chip cookies. At scale, and for industrial purposes, Search is one of the tools to help optimize supply chains. Search algorithms are a useful category of AI, but they treat problem-solving holistically and without accounting for constraints.

A search algorithm can be best pictured by thinking through a grid environment. In a grid, a search algorithm takes your initial position (such as 0,0 or the origin on a coordinate grid), applies the possible set of actions you can take, such as rules for chess piece maneuvers, calculates what will happen after each action, weighs them as possible solutions to achieving a goal state (such as arriving at 10,0), and weighs the actions across the specified action cost function, such as each move costing 1 point. Good search algorithms help an actor to find a low-cost solution using the least amount of resources, including the cost of calculating complex solutions.[24]

For example, when a computer is taught to attempt early computer games, such as Pac-Man, it can be taught to achieve a high score by finding each dot on the way to the piece of fruit. But, searching to achieve a relatively simple goal like "collect cherries in Pac-Man" can be made more difficult by increasing the complexity of the rules of the game, such as the need to avoid dead-ends and ghosts in the mazes. To teach a computer to fully play Pac-Man — where it not only collects fruits but also avoids dead-ends and ghosts — requires an additional form of symbolic AI, Planning and Scheduling.



ii. Planning and Scheduling

For more complex sets of variables that include constraints, such as finding a route to a city airport without using highway, or by a particular method such as train, an AI will normally be programmed by combining Search algorithms with Planning and Scheduling programming.

Planning algorithms are advanced Search algorithms that treat problem solving more like the way a human would. These planning algorithms can account for a variety of situational constraints, such as incomplete information, time constraints, and non-redundancy of resources. These types of models are used to design work plans, strategic design, and logistics planning for tasks like helping the Hubble Space Telescope arrange its scans of

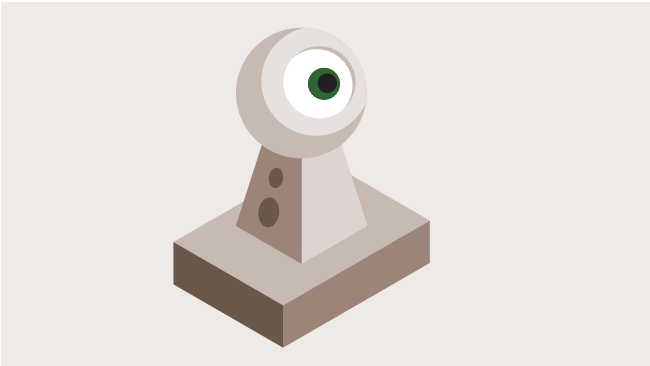
the universe.[25] When this sort of program includes time components, they also include Scheduling algorithms to show which steps are dependent on other factors, and which may take place simultaneously versus those that must occur sequentially. Global supply chain management, metropolitan transportation planning, and energy distribution grids all use these forms of Symbolic AI.

iii. Expert Systems

Some of the first Symbolic AI systems helped experienced practitioners in various professional domains to make, or teach others to make, complicated decisions, thus giving them the name "expert systems." **Expert systems** identify possible solutions by combing through and combining multiple types and layers of information. Expert systems sift through information using the reasoning and logic common to a particular profession or specialty, such as medicine or engineering.

The key components of expert systems are an extensive, detailed "knowledge base," the "inference engine," and the available working memory. The knowledge base is the network of rules, logical statements, and domain specific knowledge and reasoning that defines the expertise of humans. The knowledge base is an engineered product constructed to organize the rules, norms, protocols, and standards of the specialty, including ranking and ordering them in sequence of the information needed. This organization is done using a system of symbols comprehensible to both humans and computers such as programming languages, dictionaries for natural language processing, or established rules for maneuvering around a defined environment, such as a chess board.

Thus, expert Systems augment the decision-making capacity of professionals in a specific topical domain by systematizing their most expert levels of knowledge into heuristics (rules) and designing systems that can apply them to new situations or large sets of data inputs. These systems are also the basis for other forms of algorithmic reasoning, such as some forms of natural language processing, and can be combined with scheduling and planning systems, for use cases like robotics, and even in newly conceived situations where the solution to a problem or the best action to take is not immediately obvious even to the human observer.

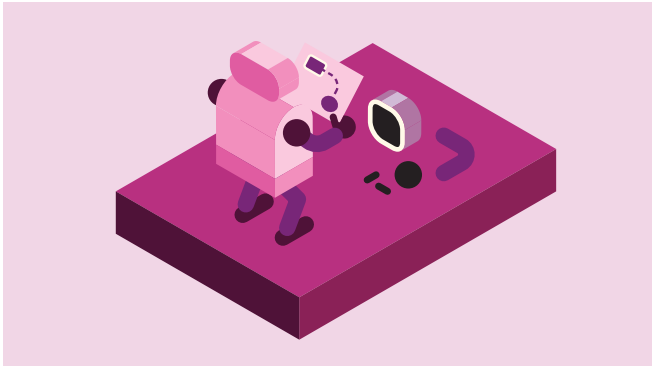


C. Computer Sensing

Perhaps the most intriguing use of AI and machine learning today is the ability to design and train computers to "sense" and then evaluate the physical environment around them. Computers can be designed with a range of sensors that simulate seeing, hearing, smelling, even tasting, and thus can be trained to collect, process, and respond to active or passive stimuli extending well beyond that of human sensory perception in both scope and accuracy. These sensors might perceive and measure light (including infrared spectrums), touch, distance (range), temperature, acceleration, and speed. Although early programmers assumed that having computers interpret their environments visually would be one of the easiest tasks, creating the algorithms essential for computer vision has proven to be an extraordinarily difficult challenge that now routinely combines aspects of rules-based and symbolic AI together with machine learning.

Furthermore, computers are being trained to "hear," "smell," or even "feel." Differentiating and classifying sounds is taught to computers in much the same way that "seeing" is. A similar set of algorithms, combined with neural networks (discussed under Machine Learning, below) are designed to analyze and classify sounds. Teaching computers to "smell" means to classify the molecules detected, and makes use of older forms of AI, combinations of AI and machine learning, and more recently, a newly designed form of neural networks called "graph neural networks."

Beyond the physical senses, there is research designed to train computers to both perceive and demonstrate emotive behavior from or on par with humans. However, attempting to design models which can reliably identify, mimic, or initiate emotive behavior remains in the early stages, is currently imprecise and inconsistent, and even where accurate, triggers some of the most challenging ethical and social questions.

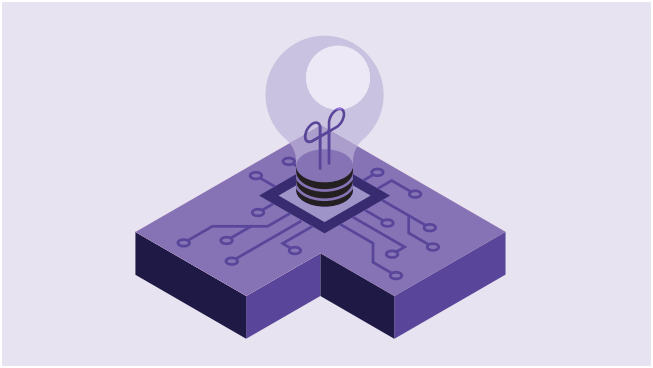


D. Robotics

Robotics is a field at the intersection of physics, engineering and computer technology that produces machines that substitute for, augment, or replicate physical human actions. They do not all, or even most, look humanoid. But in all their varied forms robots are gaining intellectual and mechanical capabilities that are largely due to the continued expansion of the AI systems powering them. They are built on Rules-based programming, and likely incorporate Sensing in many variations, and in the more complex systems, incorporate multiple Machine Learning algorithms.

Like General AI, many of our ideas about robots come from science fiction, and may mean different things to different people. However, robotics in an AI context is likely to be a much less comprehensive machine. The working definition usually includes some sort of mechanical device, with the specific abilities necessary to complete a physical task, in a particular environment, with specified parameters. Robots require a power source of some kind and contain varying amounts of programming — there are plenty of robots with no AI involved, but more and more are incorporating at least some level of advanced model or system. Some are designed to operate independently, or in many cases, to carry out tasks while in contact with cloud-based computational resources, or carry out specialty operations under a human's direct control.

The scope of robotics applications is expanding quickly. Twenty years ago, most robots were doing things like assembling cars in factories. These consisted mainly of mechanical arms tasked with routine, repeatable tasks for attaching or manipulating car parts. But by today, self-operating machines and self-propelled units explore extreme climates on Earth and other planets, assist warfighters and law enforcement, and are increasingly used in many aspects of healthcare and hospitality services.



E. Knowledge Engineering

Knowledge Engineering is a field of AI oriented to build systems that emulate the judgment and behavior of a human expert by codifying knowledge as rules and relationships between data. These systems represent knowledge as directed acyclic graphs which are able to express complex calculations and logical eligibility rules. The graphs can be easily queried and and the results reasoned to automatically produce a calculation or decision result. When reasoning using a Knowledge Engineering system, a backward chaining algorithm is typically used. Backward chaining starts from the goal and works backward to determine what facts must be asserted so that the goal can be achieved.

As an example, Knowledge Engineering powers the programs designed to provide individual and business users with the ability to comply with the ever-changing taxation rules and regulations. Because there are so many and they change so frequently to greater and lesser degrees, it would be nearly impossible for an individual to effectively identify, extract, and reconcile the new rules against the old ones. To manage this at scale for people and businesses generally, an ensemble of rule-based algorithmic approaches are adopted: Natural Language Processing is used to review current laws and extract pertinent information; graph algorithms, such as networked analysis, can show the relationship of new rules to previous instances of the rule and also reflect the impact of other applicable rules. The information extracted can be further processed into ontologies (representations of abstract concepts) that establish the terms to encode for use in subsequent applications, such as those which forecast taxable income streams and associated revenue for states and the federal government. And thus the whole is created to guide a user through a detailed but highly individualized process based on their own information, against the background of the most current rules.

F. Natural Language Processing

Designing computers to speak, as well as to understand speech and text, is one of the most fundamental functions necessary for the general progression of AI systems, and research on these various capabilities all involves some forms of natural language processing (NLP) — a combination of Symbolic AI and Machine Learning. Because language is itself a set of rules (grammar, syntax, verb forms, etc.), individualized language programs can be organized to process data according to those specific rules to classify or predict language. For example, when a large body of text (corpus) is examined using a dictionary (a set of search terms of interest) or parts of speech are tagged, the rules are used to construct the outputs, whether that be restatements, translations, or a sentiment analysis score. NLP covers the full scope of systems designed to understand text, perform analysis or translation, and interpret and create text, as well as understanding spoken inputs, and speaking in return.

G. Risks and Benefits – Rules-based AI

Rules-based or Symbolic AI may seem simple or even old-fashioned compared to its "smart," machine learning counterparts. However, many apps commonly used today, as well as much of the AI used to power the contemporary economy, make use of these variations of AI. The power of these systems is in solving design challenges for integrated circuits such as on computer chips, creating spam filters for emails, and crafting workday schedules. They can be applied to exceedingly complex problems. Most recently, for example, research into protein-folding structures, such as AlphaFold, build on the earlier work of assembly-line designs similar to the robotic assembly of machine parts, are based on rules-based AI.[26]

However, as useful as symbolic systems are, they do have some limitations. For example, search and planning-driven navigation systems may be unable to optimize for the fastest route in a way that accounts for all user preferences, such as external safety considerations or for roads with a history of flooding. And like any AI system, adverse consequences may arise from flaws in the model components or the data, whether from errors in the knowledge base or inaccuracies within the inference engine for knowledge engineering.

These systems' highest value is most often as they augment humans already carrying out these processes. Expert systems facilitate more accurate and faster diagnoses and some treatment processes but have not replaced the need for human doctors. While domain knowledge can be carefully engineered, the process of maintaining and expanding a knowledge base to account for new information introduces complexity, overlap, and the potential for errors within computerized expert systems. In the complex arena of medicine and diagnostics, keeping a knowledge base current presents an ongoing challenge, with the potential risk that patients will not receive the latest, best, or most effective treatments.

There are also considerations like risks to physical safety when general planning algorithms are not appropriately tooled by human designers, such as those used to direct industrial robots or other automated industrial processes. Likewise, planning algorithms such as those which might make up the logistics strategy for delivering pandemic vaccines or other crisis response materials, may not provide optimized plans if they are not written to sufficiently account for all the challenges involved, such as reaching remote areas, prioritizing population cohorts for tiered receipt, and storage and expiration — some of which may be difficult to define in new contexts, or generate requirements for which information is not available. As with all automated processing, the quality of the outputs will always be driven by the quality of the design and the availability and accuracy of data inputs behind the model or system.

The newest designs of algorithmic systems are the various forms of Machine Learning (ML). As mentioned, the attention and public focus on these systems as they have begun to pervade everyday devices and systems has been so prevalent in recent years, that for many people, ML and AI are treated as one and the same. With the advent of mobile devices such as smart phones and tablets, machine learning has been implemented to provide everything from weather apps to video and book recommendations, personalized healthcare and fitness platforms, and programs to accelerate and support employment and educational needs. The average consumer may enjoy using these applications but will likely struggle to describe what machine learning is, how it works, or where it contributes to the applications found in their phones, televisions, or even appliances.

VI. AI – Machine Learning

Machine learning is distinguished from Rules-based AI in that it is coded so that the machine learning system can adjust its manipulation of the data to improve on human designated metrics for accuracy and fit, as in the case of making better recommendations. The systems identify patterns in a dataset, edit inferential rules based on those patterns and connections, and then judge the fitness of that model's outputs against the requirements set by the human programmers. When we say that a machine "learns," it means the program iterates the process enough times to perform better and better, to make the model outputs more accurate against a previously set standard for success.

Machine learning has a number of sub-categories as well. ML systems can identify patterns in data we already know something about, such as when it classifies or predicts based on similarities of existing preferences, and can also derive conclusions from data we know nothing about yet, such as when it clusters items based on unseen connections or even generates new data. While there are general methods for training machines and designing machine learning systems, many programs are ultimately unique to context, designed for a specific problem. These programs predict, classify, categorize, mine, and learn from the data. Even machine learning program that start out identically become unique because each has "learned" through its own data-driven experiences, influenced by the initial weight and importance values assigned by the programmers, and then modified over time with real data.

The primary types of machine learning training are *supervised* and *unsupervised* learning, along with variations like semi-supervised, self-supervised, multi-task, and transfer learning. Without delving into detail on each, the types are essentially defined by what type(s) of data are used to develop, train, or test the system. Supervised learning is a system using data that has been labeled by humans (a process that can raise ethical issues all on its own).[27] Unsupervised learning uses unlabeled data and exploits the connections that the computer identifies, largely independently.[28] Semi-supervised learning is in between these, when the system uses a small set of labeled examples and learns from those while also evaluating and analyzing the unlabeled parts of the same dataset.

Just as humans can learn to perform new tasks simultaneously, machines can be taught how to "multi-task" or to perform related tasks simultaneously.[29] And transfer learning is when the functional skills of a machine learning model are transferred from one domain, or use case/application, to another, with appropriate adjustments tailored to the new topic and problem. This may be done for cost savings or other efficiencies, but can result in inaccurate or even ethically problematic outcomes if the updates made for the new use were not sufficient to account for different applications, and any associated risks.[30] Nevertheless, transfer learning is increasingly common as companies seek to maximize their return on investment. Particularly given the environmental costs associated with developing new or state of the art machine learning algorithms, transfer learning is increasingly being explored.[31]

Machine Learning is applied in areas like healthcare, retail, e-commerce, recommendation engines, self-driving cars, online video streaming, IoT (connected devices, voice assistants, connected home appliances), and transportation and logistics, as well as many, many others.

A. Reinforcement Learning

Some machine learning algorithms are designed with a model of learning that is based on exploration and trial-and-error without explicitly relying on existing data inputs. These systems seek to reach a particular level of accuracy by generating various outputs, checking them against expectations, and then continuously editing their process to get closer and closer to a match between solutions discovered and goals. Various factors of the outputs — such as money, time, or other resources — are given weighted values, and the algorithm is then tuned to explore in such a way that gains it the maximum reward by performing those actions that bring the highest reward values with the least errors. In other forms of reinforcement learning, the computer proceeds without any prior knowledge of the environment, adding each new learned fact after each action and thus building its own model of choices that lead to rewards.[32]

For example, a program designed to play a video game may be designed where the reward is gold coins and the errors are the loss of game lives. A reinforcement learning system will figure out how to maximize the number of coins gathered while also reducing the number of lives lost in a video game by repetitively exploring through every possible set of choices, even without preset guidance as to what its game behavior should include.

Reinforcement learning algorithms are what drive some of the notable forms of machine learning, such as when machines learn to play games, operate a flight simulator, or power a robot vacuum through a new environment.[33] And because reinforcement learning does not require an existing large set of data on which to train or operate, it can avoid some of the technical and ethical challenges involved with creating and maintaining such datasets.

B. Neural Networks

Neural Networks (NN) are a type of machine learning inspired by the human brain. They differ from other forms of machine learning in that they are non-linear by design.[34] A NN consists of a web of interconnected entities known as nodes; each node carries out a simple computation, similarly to the neurons in the human brain. Neural Networks are a collection of the algorithms used in Machine Learning for data modelling operating within this graph of nodes. Data passes through several layers of interconnected nodes, as each node classifies the characteristics and information of the previous layer before passing the results on to other nodes in subsequent layer. The layers of networks pass the data through hierarchies of various concepts, which, like other machine learning models, allows them to learn through evaluating their own errors.

The first, or input layer, is where data is received, such as data uploaded to a cloud service or live sensor data, and is then subject to a mathematical function that manipulates the data for further use over multiple iterations. The input layer to a learning system is not simply data ingestion but is an active path of calculation. The input layer leads to (at least one) hidden layer, and finally to an output layer. Each layer contains one or more nodes. By increasing the number or complexity of the hidden layers, you increase the computational and problem-solving abilities (as well as the computational costs required).[35] Deep Learning (discussed below) is defined as NNs with more than three layers.[36] Neural Network-based systems can address business challenges such as sales forecasting, consumer research, risk management predictions, and character recognition, among other things.

Neural Networks have multiple variations, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Another form, Generative Adversarial Networks (GANs) is discussed separately, below.

Convolutional Neural Networks (CNNs) are uniquely suited to operate on image data and are the algorithmic forms behind many facial recognition systems. CNNs use "kernel" filters in the input layer, meaning the image data is filtered in repeated and overlapping sections and then fed forward through the remainder of the system as a highly refined "feature map" of the image. Most CNNs are feed-forward systems which learn to see the features of an image by applying multiple filters in parallel, something like viewing the same image in up to 512 ways for a complete mental model of the image. CNNs "see" images somewhat like human artists see an object or a body — in ways that consider perspective, depth, and lighting, and how these affect the representation.

For this reason, CNNs are used in objects detection and classification systems. CNNS can detect and differentiate discrete objects within an image, such as road signs, or cats. They commonly operate by finding the "edges" of different objects, and then comparing the arrangement of edges, with other images they've previously identified.[37] (When this doesn't work as accurately as desired, undulating sand dunes may be mistaken for a reclining human body, for example.) CNNs can compare objects or faces in a 1:1 or 1:many fashion, depending on the design of the system. In other words, if the system is trying to find all the traffic lights in the image, then every item is simply "yes, traffic light," or "no, not traffic light." Other systems may want to identify as many items in the image as possible, without knowing in advance what they might be.

In contrast, Recurrent Neural Networks (RNNs) learn from data where timing and sequencing are important features, to answer questions around forecasting — predicting changes in air quality, the next word in a sentence, or stock market risk management estimates based on numerous highly variable factors. RNNs begin with data from the input layer passed through the hidden layers to the output layers, however, they include loops within the hidden layers that mimic a type of short-term memory of what has already been processed. RNNs can also map a single input to multiple forms of output. This means that RNNs are well suited for problems such as translation and voice classification where one input — a letter or a phoneme (smallest piece that distinguishes a word) — can lead to different outputs.[38]

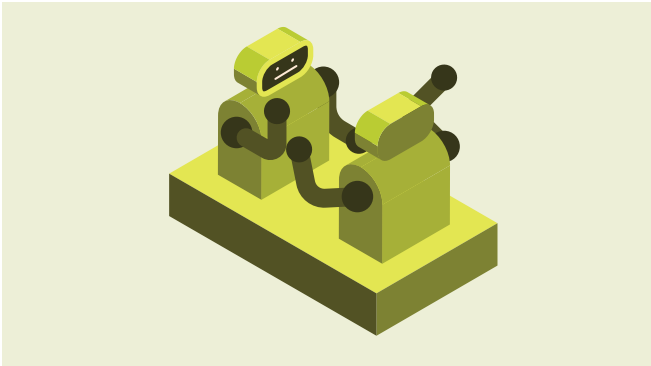
VI. AI — Machine Learning (continued)

C. Deep Learning

A subset of machine learning, Deep Learning simply refers to the number of total connections a program makes in between the input layers and output layers, in that it has more than three layers. Between input and output layers are "hidden" layers which perform additional calculations through refining the weights applied to the various features of the data. The output layer generates the final result to the user, usually a "recommendation," along with the confidence level of the result ("yes, that animal is a cat").

Deep learning models, particularly when applied to CNNs or RNNs, comprise exceptionally powerful systems for complex tasks such as language translation that could not be accomplished by rules-based AI, or by linear machine learning where there is only one computational pass of the data from input to output. The "looping" feature of deep learning models allow them to use both forward and backward computational paths through the layers, making it possible to identify obscure or highly nuanced patterns and relationships in the data.[39]

In addition to editing its own inferential reasoning pathways, Deep Learning program systems can create entirely new, additional hidden layers between the input layer and output layer without human intervention. This is part of what makes "transparency" or "explainability" so challenging for the users, or even designers, of a deep learning program. The path through the computational layers may be different even for consecutive data inputs, and the structure of the network processing is changing and evolving in real time. Following exactly what happened (what computations were applied, in what sequence, and with what factors or features involved) can be difficult to observe or recreate. However, there is ongoing research on explaining deep learning systems, attempting to make progress toward seeing into the "black box" to understand, describe, or replicate what the systems did to reach a particular conclusion.[40]



D. Generative Adversarial Networks

Generative Adversarial Networks (GANs) are the newest variation of ML being developed, and are based on a pair of neural networks that learn by attempting to get the better of each other. First, the "generator" of the pair creates an output (e.g., an image) based upon the initial human programming. The other network, the "discriminator," has been programmed with criteria to evaluate the output (e.g., what the image should look like). The discriminator considers the output, and critiques it — essentially determining whether it is real, or correct. Early in the cycle, the initial outputs are likely to be far off from what is desired. The discriminator's feedback is then incorporated into the program and the generator continues to churn out results, and the feedback loop continues, until the generator produces data that the discriminator believes meets the quality expectations. In the case of an image, this is the generator "fooling" it into thinking a generated output is real. GANs are very new and their capabilities are still being explored, but they are the systems used, so far, to produce "deep fakes," contemporary works of art, music, or writing in the style of long dead masters, or to create entirely unique compositions by the AI system.

E. Risks and Benefits — Machine Learning

Machine learning programs are only as good as their design and training. Just as humans risk injury when they rush to run great distances or perform complex sports maneuvers without proper preparation, machine models risk generating faulty outputs, or even causing harm if they are not well-designed and if their training is not continuously monitored and updated. That these systems "learn" without additional human programming does not mean they should operate without human oversight. If a system initially learns from a limited, badly organized, or insufficiently representative data set — whether the limits are due to the size of the data, the variety of data, the completeness of the data, or the veracity of the data — it cannot work properly "in the wild."

Machines learn under guidance, application, and evaluation by humans. If machines are applied to problems that do not truly match their design, then they will deliver poor results. If a machine is directed to find patterns in data that is too limited to fully represent the knowledge the computer needs to solve the problem presented, the machine learning program will be underfit to the problem.[41] One example of the underfitting problem comes from uses of machine learning for predicting complex disease dynamics. If a machine is asked to identify the major causal drivers for a disease, which nurses and therapists know is strongly influenced by social factors but the machine is only given laboratory test data to learn from, then it will be underfit for the purpose of helping medical teams manage the disease. Likewise, if a similar system is only trained on data from a single hospital system that uses a single set of vendors to provide health records and reports of diagnostic tests, the learning system may be underfit when applied to another hospital system using another vendor›s records.

As has been covered extensively in the media and academia, machines that replicate historical human-run systems, and which are therefore trained on the historical data from those systems, are assuredly going to perpetuate any biases or inequities of those systems. For example, predictive models such as those that are used by financial organizations to predict loan risk, if trained solely on historical loan data, will exhibit the same historic biases tied to race, educational institutions, or other discriminatory factors. That is, some customers will be privileged in their receipt of loans while others are systematically excluded. In fact, in such systems, the biases can be perpetuated even when the programming has attempted to minimize such aspects, because the systems are capable of such nuanced associations (e.g. when zip codes become proxies for socio-economic status or race).

These bias challenges can sometimes be mitigated by addressing data quality, sufficiency and representativeness, but the biases are not always easy to identify and isolate.[42] Alternatives to improve these models include designing systems using reinforcement learning (without the need for historical data), or synthetic data sets, but these are also not immune to the inherent bias carried over by programmers and social structures and must be evaluated and monitored carefully.

While reinforcement learning models do not use historical data, they nevertheless sometimes produce undesirable outputs if their goals are not carefully designed and defined. For example, during the trials for fighter pilot systems, viewers watched reinforcement learning algorithms battle one another by spinning in loops at forces impossible for human pilots to withstand or by zooming towards the sky or the ground to avoid one another.[43] The system had been designed to maximize points based on avoiding being shot down, but the limitations of human bodies had not been included in the code. Thus, if put into a real environment, such systems would cause severe harm to a pilot and costly equipment. Likewise, reinforcement learning systems must have the "common sense" programming to set general parameters as well. Notably, some systems of security robots, which are hard-coded to avoid confrontational humans, sought escape by diving into a nearby fountain.[44]

When CNNs are applied to image models, they can be highly successful, reliably recognizing not only that a dog is not a cat but that a Labrador retriever is not a Pekingese. However, if poorly trained, they will distinguish wolves from dogs simply on the basis of snow in the picture, or mis-categorize men and women of particular races, or with specific health conditions. When moving beyond image recognition to individual identification systems such as facial recognition, models have been repeatedly shown to fail badly when the training datasets are insufficiently diverse and representative of the faces that will be encountered once the system goes into production.[45] There are certainly facial recognition systems now that have corrected this problem and are highly accurate across all demographics without significant variance, but these systems remain in the minority (and are the most expensive) of those available for commercial use. Facial analysis systems that attempt to categorize people by gender, or identify characteristics or emotions remain highly flawed for technical accuracy as well as ethically suspect.

RNNs face the biggest challenges related to transparency, as discussed above, as they have an element of inscrutability to their structure that makes them difficult to explain or replicate. Explaining fully how the loops between layers in RNNs change the data and thus impact the output score remains an elusive goal for explainable AI. While it may not be of immediate importance to know why an RNN offers one translation of a phrase or another, it certainly is important to know that these neural network outputs suffer from fallibilities that mean they can make inaccurate or inappropriate translations which humans should be cautious to trust.[46]

This delineation of the risks of various aspects of ML should not be read as outweighing the many benefits, conveniences, cost savings, and efficiencies that have already resulted from these systems, with more being considered and developed all the time. From expanding access to credit and financial systems, improving health care, supporting educational opportunities, and the entire infrastructure of connected devices and home services, ML has improved our lives in many ways. These gains will only continue to grow. Specifying the risks simply outlines the detailed and critical responsibility incumbent on those who develop, sell, and employ such systems to use (and refrain from using) them responsibly, in alignment with ethical values and a prioritization of human well-being.

VII. Conclusion

AI is a field of science that encompasses technical, social, and policy considerations. As with any technology, there is no "neutral" system — the choices made of what to automate; how to determine the type, style, and features included; sourcing the data; and applying the system in specific contexts are all design choices that carry implications for equity and harm. Automating human functions and behaviors carries the inherent risks of automating human errors and shortcomings.

However, AI systems can also add specific and extended value to many fields, providing efficiencies in cost and time, as well as accuracy and reliability. Mobile services that provide banking, communications, and safer, smoother travel, among others, would not be possible without it. Low cost goods, access to education and employment, and global efforts toward environmental health and sustainable practices are all made feasible or significantly more effective by AI. Healthcare in particular has been entirely disrupted and generally improved by the capabilities of these systems.

As individuals, private companies, governments and regulators seek to safely and responsibly implement these programs into the everyday lives of individuals and infrastructures, they must understand the technology as it functions within the particular industry or application, and carefully consider where and how appropriate restrictions should apply. This paper will hopefully make that task easier, provide an understanding of the complexities, the opportunities, and the limitations of these programs so that they can be an overall benefit and boon to humanity.

Endnotes

1 Tanha Talaviya, Dhara Shah, Nivedita Patel, Hiteshri Yagnik, Manan Shah. 202. Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides, Artificial Intelligence in Agriculture, Volume 4, Pages 58-73. Available at: https://doi.org/10.1016/j.aiia.2020.04.002.

2 National Highway Traffic Safety Administration. 2017. "Automated Driving Systems: A vision for safety". Available at: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf

3 Food and Drug Administration. 2021. "Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

4 An important cultural note must be pointed out before examination of these foundational disciplines: Just as artificial intelligence does not belong to one culture or computer programming to one language, philosophy and mathematics do not exist only in cultures with a touch to the Northern hemisphere, Mediterranean or Atlantic Oceans, or Abrahamic religious traditions. Philosophy is a form of cultural expression common to cultures striving to answer the questions: where did we come from, what are we, and how do we know? Mathematics is both the language of any market but also the language of abstract expression of natural constants. The brutal realities of history and material preservation mean that the set of knowledge recognized as philosophy and mathematics includes disproportionately more of the thinking traditions written rather than spoken or performed, written on paper rather than on stone, and composed in alphabetic and promiscuous languages that translate words easily, even at risk of the loss of meaning. Due to the geographical and historic situation of the authors of this report we rely on the written word, Western cannons of thinking, and the English language primarily.

5 Jordan, Sara R., and Phillip W. Gray. *The ethics of public administration: The challenges of global governance*. Baylor University Press, 2011; Jordan, Sara R. "The innovation imperative: An analysis of the ethics of the imperative to innovate in public sector service delivery." *Public Management Review* 16, no. 1 (2014): 67-89.

6 MacIntyre, Alasdair. *A Short History of Ethics: a history of moral philosophy from the Homeric age to the 20th century*. Routledge, 2003; Blackburn, Simon. *Being good: A short introduction to ethics*. OUP Oxford, 2002.

7 Schuelke-Leech, Beth-Anne, Sara R. Jordan, and Betsy Barry. "Regulating Autonomy: An Assessment of Policy Language for Highly Automated Vehicles." *Review of Policy Research* 36, no. 4 (2019): 547-579 Jordan, Sara R., and Phillip W. Gray. "Clarifying the concept of the "Social" in risk assessments for human subjects research." *Accountability in research* 25, no. 1 (2018): 1-20.

8 Priest, G. (2017). *Logic: A very short introduction* (Vol. 29). Oxford University Press; Floridi, Luciano. *Information: A very short introduction*. OUP Oxford, 2010.

9 Apt, Krzysztof R. "Logic Programming." *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics (B)* 1990 (1990): 493-574.

10 Deisenroth, Marc Peter, A Also Faisal and Cheng Soon Ong. *Mathematics for Machine Learning.* Cambridge University Press. (2020); Wilmott, Paul. *Machine Learning: An Applied Mathematics Introduction.* Panda Ohana. (2019).

11 Hogg, Tad, and Bernardo A. Huberman. "Artificial intelligence and large scale computation: A physics perspective." *Physics Reports* 156, no. 5 (1987): 227-310; Haugeland, John. *Artificial intelligence: The very idea*. MIT press, 1989.

12 Inside Big Data. 2020. "The 6 types of data everybody should know to avoid confusion". Available at: https://insidebigdata.com/2020/12/25/the-6-types-of-data-everybody-should-know-to-avoid-confusion/

13 Harford, Tim. "Big data: A big mistake?." *Significance* 11, no. 5 (2014): 14-19; Neef, Dale. *Digital exhaust: what everyone should know about big data, digitization and digitally driven innovation*. Pearson Education, 2014; Cunningham, McKay. "Next generation privacy: the internet of things, data exhaust, and reforming regulation by risk of harm." *Groningen Journal of International Law* 2 (2014).

14 Khan, M. Ali-ud-din, Muhammad Fahim Uddin, and Navaram Gupta. 2014. "Seven V's of Big Data: Understanding big data to extract value". Available at: http://www.asee.org/documents/zones/zone1/2014/Professional/PDFs/113.pdf

15 Halpern, Orit. *Beautiful data: A history of vision and reason since 1945*. Duke University Press, 2015.

16 Mason, Richard O. "Four ethical issues of the information age." *MIS quarterly* (1986): 5-12; https://www.talend.com/resources/what-is-data-processing/

17 Dangeti, Pratap. *Statistics for machine learning*. Packt Publishing Ltd, 2017.

18 Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. "Points of significance: statistics versus machine learning." (2018): 233.

19 Sciforce. 2019. "AI hardware and the battle for more computational power". Available at https://medium.com/sciforce/ai-hardware-and-the-battle-for-more-computational-power-3272045160a6

20 Apple, Inc. 2021. "Apple design resources". Available at: https://developer.apple.com/design/resources/

21 An example of an algorithm for a specific domain represented as rules for a human or computer to follow includes the Mayo Clinic urinalysis adulterant survey algorithm, which is available at: https://www.mayocliniclabs.com/it-mmfiles/Adulterant_Survey_Algorithm.pdf

22 Symbolic and neural systems are not walled off from one other. Bader and Hitzler (2005) show how these two can be brought together. Bader, Sebastian and Pascal Hitzler. 2005. "Dimensions of neural-symbolic integration—A structured survey". Available at: https://arxiv.org/pdf/cs/0511042.pdf

23 PCMag. 2021. "Inference engine". Available at: https://www.pcmag.com/encyclopedia/term/inference-engine

24 Completeness is when an algorithm will find a solution when there is one or will report failure when a solution cannot be identified. Optimality is identification of the lowest cost solution among all options. Complexity is the amount of time or memory space needed to perform a search.

25 Samson, Roberto J. 1998. "Greedy search algorithm used in the automated scheduling of Hubble Space Telescope activities". Proceedings of SPIE 3340, Observatory Operations to Optimize Scientific Return10.1117.12.316498.

26 Callaway, Ewan. 2020. "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures". Available at: https://www.nature.com/articles/d41586-020-03348-4

27 Savage, Saiph. 2020. "AI needs to face up to its invisible-worker problem". MIT Technology Review. Available at: https://www.technologyreview.com/2020/12/11/1014081/ai-machine-learning-crowd-gig-worker-problem-amazon-mechanical-turk/

28 Raw word clouds, which are representations of term frequency (word count) and where no terms have been removed or no color coding applied, are a visual representation of unsupervised learning. These are not always particularly helpful and the frequency of seeing word clouds seems to be on the wane. M. Hearst, E. Pedersen, L. P. Patil, E. Lee, P. Laskowski and S. Franconeri, 2019. "An Evaluation of Semantically Grouped Word Cloud Designs," in IEEE Transactions on Visualization and Computer Graphics. http://dx.doi.org/10.1109/TVCG.2019.2904683.

29 Xu, Yichong, Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. "Multi-task learning with sample re-weighting for machine reading comprehension." *arXiv preprint arXiv:1809.06963* (2018).

30 Jamshidi, Pooyan & Velez, Miguel & Kästner, Christian & Siegmund, Norbert & Kawthekar, Prasad. 2017. Transfer Learning for Improving Model Predictions in Highly Configurable Software. 10.1109/SEAMS.2017.11.

31 Sharir, Or, Barak Peleg, and Yoav Shoham. 2020, "The Cost of Training NLP Models: A Concise Overview." *arXiv preprint arXiv:2004.08900*.

32 Bhatt, Schweta. 2018. "5 things you need to know about reinforcement learning". Available at:. https://www.kdnuggets.com/2018/03/5-things-reinforcement-learning.html

33 Brandon, John. 2016. "Why the iRobot Roomba 980 is the greatest lesson on the state of AI". Available at: https://venturebeat.com/2016/11/03/why-the-irobot-roomba-980-is-a-great-lesson-on-the-state-of-ai/

34 Willems, Heather. 2016. "The power of non-linear thinking". Available at: https://www.americanexpress.com/en-us/business/trends-and-insights/articles/power-non-linear-thinking/

35 B.K. Lavine, T.R. Blank, 2009. "3.18 - Feed-Forward Neural Networks", Editor(s): Steven D. Brown, Romá Tauler, Beata Walczak, Comprehensive Chemometrics, Elsevier, Pages 571-586.

36 Brownlee, Jason. 2020. "Crash course on multi-later perceptron neural networks". Available at: https://machinelearningmastery.com/neural-networks-crash-course/

37 Ziou, Djemel, and Salvatore Tabbone. 1998. "Edge detection techniques-an overview." *Pattern Recognition and Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii* 8: 537-559.

38 For language processing tasks, attention models, such as transformers, are now used in the most common and best performing language systems such as "BERT" (Bidirectional Encoder Representations from Transformers) or its permutations RoBERTa, DistilBERT, BioBERT, or BEHRT. While RNNs learn to identify patterns by observing sequences as dependent on time or sentence structure, meaning from start to finish, Transformers do not need to process data sequentially to identify sequences as part of the output of a transformer model. This means that transformers are able to process data in parallel and without some of the problems of RNNs. In layman's terms, transformers can interpret the whole book by reading all of it at once, rather than page by page or chapter by chapter, relying on a mental model of time or space development such as how humans learn a story by watching the chapters unfold. Transformers can absorb all of the information without such a mental model. See for more: Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A robustly optimized bert pretraining approach." *Available at: https://arxiv.org/abs/1907.11692; or* Staliūnaitė, Ieva, and Ignacio Iacobacci. 2020. "Compositional and Lexical Semantics in RoBERTa, BERT and DistilBERT: A Case Study on CoQA." Available at: https://arxiv.org/abs/2009.08257

39 A very simple analogy to describe the difference between an input layer and loading data is the difference between watching a 3D film with the glasses and without. The dazzling 3D features appear with the red-blue filter of the glasses but do not when the glasses are not worn.

40 Montavon G., Binder A., Lapuschkin S., Samek W., Müller KR. (2019) Layer-Wise Relevance Propagation: An Overview. In: Samek W., Montavon G., Vedaldi A., Hansen L., Müller KR. (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol 11700. Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_10

41 Jabbar, Haider K. and Rafiqul Z. Khan. 2015. "Methods to avoid over-fitting and under-fitting in supervised machine learning". Available at: https://www.researchgate.net/profile/Haider_Allamy/publication/295198699_METHODS_TO_AVOID_OVER-FITTING_AND_UNDER-FITTING_IN_SUPERVISED_MACHINE_LEARNING_COMPARATIVE_STUDY/links/56c8253f08aee3cee53a3707.pdf

42 Lardinois, Frederic. 2017. "Intuit launches QuickBooks Capital, a small business lending service powered by AI". Available at: https://techcrunch.com/2017/11/07/intuit-launches-quickbooks-capital-a-small-business-lending-service-powered-by-ai/

43 Defense Advanced Research Projects Agency. 2020. "AlphaDogFight Go Virtual for Final Event". Available at: https://www.darpa.mil/news-events/2020-08-07

44 Garun, Natt. 2017. "DC security robot quits job by drowning itself in a fountain". Available at: https://www.theverge.com/tldr/2017/7/17/15986042/dc-security-robot-k5-falls-into-water

45 Cornejo, Jadisha Yarif Ramírez et al. "Down syndrome detection based on facial features using a geometric descriptor." *Journal of medical imaging (Bellingham, Wash.)* vol. 4,4 (2017): 044008. doi:10.1117/1.JMI.4.4.044008;

46 FPF MasterClass Webinar, "Machine Learning and Speech," December 2020, presentation by Professor Marine. Carpuat, https://www.youtube.com/watch?v=uRJsTy50Sqs

ACKNOWLEDGMENTS

Many thanks to the supporters of Future of Privacy Forum, especially members of the Artificial Intelligence and Machine Learning Working Group, including industry representatives, academic experts, and civil society advocates, who all contributed to this paper by way of suggestions, feedback, and technical details.

For further information about Future of Privacy Forum and its work on public policy regarding Artificial Intelligence and Machine Learning, please contact info@fpf.org.

The Future of Privacy Forum (FPF) is a catalyst for privacy leadership and scholarship, advancing responsible data practices in support of emerging technologies. FPF is based in Washington, DC, and includes an advisory board comprising leading figures from industry, academia, law, and advocacy groups.

1400 EYE STREET NW | SUITE 450 | WASHINGTON, DC 20005 INFO@FPF.ORG | 202-768-8950