Artificial Intelligence and the Courts:

MATERIALS FOR JUDGES

Artificial Intelligence – Foundational Issues and Glossary

SEPTEMBER 2022

AAAS
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Artificial Intelligence and the Courts: Materials for Judges

The American Association for the Advancement of Science (AAAS) is honored to have been entrusted, by the National Institute of Standards and Technology (NIST), with the task of developing educational materials on artificial intelligence (AI) for judges and courts.

AAAS therefore offers this compilation of educational materials for judges, covering a wide, yet appropriate, set of issues. (Please see the list below). AAAS' goal is to provide a set of user-friendly and accurate, yet readily comprehended, definitions, analyses, and perspectives, on a variety of terms and topics with which the judiciary ought to become familiar.

The materials contained herein were developed by teams of scientific and legal experts who focused on a particular topic. The topics considered worthy of inclusion were selected based both on the mandate provided by NIST and guidance received by AAAS from an Advisory Committee composed of a large and diverse group of legal and AI experts. Drafts of the materials were subsequently submitted to Advisory Committee members, and outside expert "Reviewers," to obtain any suggestions for adjustments before each team of authors finalized their contribution (paper, podcast, annex, etc.).

It is not expected that courts will become experts regarding these sometimes complex or technical matters. Rather, this collection presents facts and overviews in a manner intended to make judges aware of key issues and to enable courts to find useful information contained herein, easily.

Finally, it is hoped that courts will appreciate certain innovative elements of this product, notably the inclusion of podcasts. These will provide courts with facts and analysis of important questions in a format that courts may find agreeable and, given the accompanying transcripts included, useful. AAAS thanks NIST for allowing a team of experts to undertake this forward-leaning approach to providing courts with needed information and insights as part of this project.

Materials in this series include:

1. ***Artificial Intelligence – Foundational Issues and Glossary***
2. ***Artificial Intelligence and the Justice System*** **(Podcast Series and Transcripts)**
   - Episode 1: *AI and Risk Scores* (49 minutes)
   - Episode 2: *AI in the Legal Field – Commercial and Unexpected Uses* (70 minutes)
   - Episode 3: *AI, Decision-Making, and the Role of Judges* (58 minutes)
3. ***Artificial Intelligence, Trustworthiness, and Litigation***
4. ***Artificial Intelligence, Legal Research, and Judicial Analytics***
5. ***Artificial Intelligence and Bias – An Evaluation***

Acknowledgements

The American Association for the Advancement of Science (AAAS) gratefully acknowledges the authorship of Kay Firth-Butterfield, Head of Artificial Intelligence and Machine Learning and Executive Committee Member, World Economic Forum, and Karen Silverman, Chief Executive Officer and Founder, Cantellus Group. In addition, we are thankful for the contributions from Brin Elliott, Analyst, Cantellus Group.

We also extend our gratitude to the members of the Advisory Committee for their input and guidance throughout the design and production of this product: Rediet Abebe, Assistant Professor, Computer Sciences, University of California, Berkeley; Theodore F. Claypoole, Partner, Womble, Bond, Dickenson, LLP; John Cooke, Director, Federal Judicial Center; Cynthia Cwik, Senior Legal Advisor, Former Stanford Fellow, and Former Partner, Jones Day and Latham & Watkins; Herbert B. Dixon, Jr., Judge (ret.), Superior Court of the District of Columbia, and Senior Legal Advisor to the Center for Legal and Court Technology, William & Mary School of Law; Finale Doshi-Velez, Associate Professor, Computer Science, School of Engineering and Applied Sciences, Harvard University; Paul W. Grimm, Judge, US District Court for the District of Maryland; James Harris, Technology Consultant, National Center for State Courts; Joseph Sawyer, Director, Faculty Development and Online Learning, National Judicial College; Elham Tabassi, Chief of Staff, Information Technology Laboratory, National Institute of Standards and Technology; and Zachary Zarnow, Principal Court Management Consultant, National Center for State Courts.

AAAS appreciates the contributions of Solomon Assefa, Vice President, IBM Research; Nicole Clark, Chief Executive Officer, Trellis Research; Rebecca Crootof, Associate Professor of Law, University of Richmond School of Law; Patrick Huston, Brigadier General and Assistant Judge Advocate General for Military Law and Operations, US Department of Defense; Maia Levy-Daniel, Attorney and Public Policy Consultant; Matthew Stepka, Managing Partner, Machina Ventures, and Lecturer, Haas School of Business, University of California, Berkeley; Peter Stone, Professor, Department of Computer Sciences, and the Founder and Director, Learning Agents Research Group, Artificial Intelligence Laboratory, University of Texas, Austin; Rebecca Wexler, Assistant Professor of Law and the Faculty Co-Director, Berkeley Center for Law & Technology, University of California, Berkeley; and others for their review and insightful comments on previous drafts of the *Artificial Intelligence and the Courts: Materials for Judges* series.

This paper was edited by Joel Ericsen, Program Associate, and Alain Norman, Senior Program Associate, Science and the Law, AAAS. Special thanks to Theresa Harris, Program Director, Scientific Responsibility, Human Rights and Law Program, and Nathaniel Weisenberg, Program Associate, AAAS, and to Barbara “Bebe” Holloway, University of Virginia, for their contributions.

AAAS Creative and Brand Manager, Paula Fry, created the cover design for this publication.

The production of this publication and other materials in the *Artificial Intelligence and the Courts: Materials for Judges* series was funded by the United States National Institute of Standards and Technology (NIST) through Award 60NANB21D031.

Disclaimer: The opinions, findings, and conclusions or recommendations expressed in this publication do not necessarily reflect the views of the AAAS Board of Directors, its Council and membership, or NIST.

Contact: AAAS welcomes comments and questions regarding its work. Please send information, suggestions any comments to the AAAS Scientific Responsibility, Human Rights and Law Program at srhrl@aaas.org.

Cite as: Firth-Butterfield, K. and Silverman, K. "Artificial Intelligence – Foundational Issues and Glossary." *Artificial Intelligence and the Courts: Materials for Judges*. American Association for the Advancement of Science (2022). Available at https://doi.org/10.1126/aaas.adf0782.

Abstract

Foundational Issues: This paper provides a very inclusive introduction to numerous key concepts with which judges and court personnel may need to be familiar. These include: The elements and variety of artificial intelligence (AI) systems; how these may be designed, developed, or deployed; and key issues with respect to the limits and risks associated with AI. This section also includes:

- *Annex A*: "Who is Regulating AI?" In the US, and internationally, the increased use of AI is leading to various efforts to regulate it, but the AI regulatory landscape is vast, fragmented and changing very rapidly. This Annex offers a comprehensive snapshot of the current state-of-play, which courts may find useful while bearing in mind the need to monitor this evolving landscape.

Glossary: For convenient reference, this product provides, in alphabetical order, a list of important terms and words (from, e.g., "Acoustical Processing," to "GAN – General Adversarial Networks," to "Responsible AI," to "VR – Virtual Reality") and their definitions. As warranted, differences in nuances or usage are noted. Additionally, certain common abbreviations or acronyms are included.

Table of Contents

Artificial Intelligence – Foundational Issues and Glossary

Authors:

Kay Firth-Butterfield
Head of Artificial Intelligence and Executive Committee Member, World Economic Forum

Karen Silverman
Chief Executive Officer and Founder, Cantellus Group

Contributor:

Brin Elliot
Analyst, Cantellus Group

1. How Artificial Intelligence (AI) Works

Human beings are critical to the functioning of artificial intelligent (AI) systems. Those human beings, their agreements and processes will impact AI just as AI will impact humans. Therefore, for all the technological advancements, humans will remain at the center of all judicial work and philosophy.

AI represents an enormous opportunity and powerful tool for improving the efficiency and quality of any human thinking task, and it also raises challenges and critical questions (some of them existential). This is especially true from the perspective of the judiciary. The courts will encounter AI in every part of their work – including in the most profound role as being and overseeing finders of fact in a time where very real events are taking place in completely synthetic environments – so, it is imperative that we all learn to ask questions and make careful determinations about where to seek out its benefits and when to exercise caution about its risks.

2. What is AI?

Artificial intelligence[1] (AI) does not have a widely agreed upon definition. Broadly speaking, **AI is both a concept and a category of technology tools** that are powered by advanced mathematical **models** and data that can augment, replicate or improve upon the type of human cognitive task that otherwise requires thinking, beyond calculating.[2] "AI" often refers to the capability of a model to "acquire, process and apply knowledge," which is an effort to differentiate AI from other technologies that operate less "independently" to achieve their results.

[1] Bolded red terms appear in the Glossary.

[2] By a calculating task, it is meant that two inputs always produce the same outputs (e.g., 2+2 always equals 4). Where two inputs do not always produce the same outputs, that is more akin to a thinking task (e.g., going to the grocery store and putting cookies in your basket does not always mean you are going to also put milk in your basket).

Arthur C. Clark's observation that "any sufficiently advanced technology is indistinguishable from magic" would seem to fit, as within the category of AI is a wide, and ever-expanding, basket of techniques and applications that open new ways for people to see, understand and engage with the world. As with most tools, AI is **dual use**, meaning it can be put to beneficial or harmful purpose. Likewise, AI is a tool that can be made or applied well or poorly.

Most of the AI in use today is considered **Narrow AI** which refers to models aimed at performing discrete tasks and functions. In some cases, Narrow AI is already outperforming human actors.[3] In the future (and debates rage around how near or distant), we may well see **General AI** capable of autonomous existence across a full range of capabilities, with the ability to establish its own objectives, adjust to circumstances, learn and follow through on complex instructions. Currently, however, and for the near future, Narrow AI – to be sure, with increasingly generalized and sophisticated functionality – is the appropriate subject of this discussion.

For this community, which will be called upon to adjudicate disputes and interpret AI-fueled outcomes, "AI" is a deceptively uninformative term that means very little without more insight into what type of AI is being used, to do what, by whom and in what context and time frame. Similarly, as will be detailed below, the terms "**Responsible AI**," "**Ethical AI**" and "**Trustworthy AI**" are not well-defined and will likely need to adjust as the technologies and use cases evolve. Critically, the terms do not indicate human values or principles (e.g., fairness, transparency, human rights, non-discrimination, etc.) have been engineered into a particular model or process. Rather these terms will likely always reflect a mix of human process *and* technical efforts to select, design, use and monitor evermore complex AI in ways that comport with human and cultural values.

Before going into further detail about AI, two other important introductory points deserve mention: It is crucial for the judiciary, along with everyone else, to have an understanding of AI because (1) AI models and systems are increasingly deployed in commercial, academic and government settings; and (2) the **outputs**[4] of AI models used today fundamentally are **predictions**,[5] generated by correlating information and recognizing patterns from past events or instances (data) with new data to forecast the likelihood of an event or instance occurring in the future – meaning AI models offer **probabilities**[6] and carry inherent uncertainty, the degree and significance of which courts will need to evaluate.

[3] Likewise, some people refer to Augmented Intelligence to describe the narrower, contemporary function and purpose of today's technology.

[4] The result or value produced by an AI model. Outputs are generated from combining input data with the model, and fundamentally are predictions.

[5] AI model outputs that result from correlating information and recognizing patterns from past events or instances (data) with new data to forecast the likelihood of an event or instance occurring in the future – meaning AI models offer probabilities and carry inherent uncertainty.

[6] Calculations that predict the likelihood of the occurrence of a certain event.

2.1. AI's Features: Short List of Key Characteristics

Notwithstanding the variety of types and uses of AI, most AI consists of the following key elements and features:

1. A model consisting of an **algorithm**[7] that is applied to organize, evaluate and assess matches and **patterns** in **data** to produce a **prediction**, from which the model itself, or those using the model, can derive **insights**.
2. A model that is designed to **optimize** goals and insights about a particular question or objective that is **framed** and described by humans ("**goal alignment**").[8]
3. A model that can process vastly more variables, simultaneously and more quickly than any human mind can process, and so therefore can generate volumes of surprising and original insights.
4. A model that detects matches, patterns and correlations in **training datasets** of **artefactual** information or collections of prior behaviors and then applies the trained matches and patterns to new **production data** to generate predictions about new situations.
5. A model that therefore produces outputs in the form of **predictive correlations**, not proofs, about how historical data informs new events, and so always contains some magnitude of **uncertainty and error rate** (just like humans). These outputs can be predictions, recommendations or insights, or even decisions in some cases. How the outputs are used to affect outcomes is an important inquiry.
6. A model that has some **opaqueness** and cannot necessarily explain to humans how it arrives at its insights or how humans should (or should not) rely on those insights.
7. A model that – even if well-designed and intentioned – will always reflect **bias** in some respect (not always bad, but sometimes bad), because that is innate to how this technology (and humans) functions when organizing information.
8. A model that is amoral and has no innate understanding of context, ethics, or reason, or even of limits unless those limits somehow are incorporated as part of its design.
9. The opportunity for evolution at a breathtaking pace, in terms of the AI techniques in use, the data that are available and the **use cases** to which AI is applied.

AI Systems describe the ecosystem that includes AI models (themselves composed of algorithms and data), along with the humans, their organizations and any other technologies associated with their AI Lifecycle. It is important to broaden the lens to consider the *human and social systems* around the models, because a) how well those human systems operate determines how well the models and the technology really works – and the impacts they really

[7] Mathematical logic (sometimes called "rules") used to organize, evaluate and assess matches and patterns in data.

[8] Model alignment refers to the process of fitting a model to optimize goals and insights about a particular question or objective framed and described by humans.

have – in applied settings; and b) in reality, most AI today is not fully autonomous and does not fully replace any human function or task but rather, augments human function.

How well an AI system works therefore will turn on how well the **human/technology interaction** works, how well it was designed, taught, reinforced, monitored, etc. An AI model can be designed to manage a human workflow, or the other way around, for instance, and how that relationship unfolds in practice, is critical. Likewise, humans are going to be impacted, positively or negatively, by AI systems. It can be complicated to unpack how the human and the machine individually contribute to that impact.

While the terms "**Responsible AI,**" "**Ethical AI**" and "**Trustworthy AI**" are becoming more common, they likewise defy an agreed meaning. In general, these concepts are used to describe AI models and systems that are designed, built and operated with forethought and certain key attributes that protect human beings (such as fairness, safety, transparency and a respect for human autonomy) and denote a level of quality in their performance overall. At this point, and for the foreseeable future, these terms do not suggest that the models themselves are ethical or responsible; they may be engineered to defer to humans (or not) or to resolve uncertainties in one way or another, but the models themselves do not contain ethics in a human sense. On the other hand, models might be "trustworthy" in key dimensions, meaning that they are designed to be safe, reliable and verifiable, for instance.

For purposes of this discussion, it is important to acknowledge that despite the many statements of Responsible AI principles now in circulation (well over 100 of them), there is no agreed legal standard for what this means. There are existing laws that will be applied to AI and new legislation that will be passed. Community standards for what is reasonable or just adequate will also evolve.

2.2. AI in the Courthouse: Several Key Examples

Before diving into the specific issues and operating features of AI, it is helpful to briefly identify examples of where AI may show up in the courtroom today (and this list is expanding by the day), most of which will be explored further in subsequent chapters.

- **Operations:** Perhaps the easiest use cases to see are in how these tools can be used to streamline courthouse operations, from screening vast quantities of résumés to managing dockets, scheduling hearings and trials and managing jury functions and juror pools.
- **Case management**: Apart from calendaring, AI can be used for record management and retrieval, scanning huge volumes of briefing material, case law and evidence.
- **Legal research and writing**: One of the original use cases for AI includes legal research using enormous data sets of reported legal opinions. AI tools can aid in the research of historical records and practices to form predictions. In the case of analyzing and supporting sentencing recommendations, AI tools are in fairly wide use, even while criticized for their efficacy and tendency to reinforce historical bias. In light of current

capabilities, soon, courts, clerks and counsel may use AI to aid in drafting memos and opinions.

- **Discovery and evidence**: Predictive coding in discovery is another foundational use case for AI tools in the law. The full range of discovery practices and discovery disputes is likely to change and reflect the new tools and targets for discovery. What will it mean to have custody or control of information when collecting vast quantities of data? How will discovery be conducted, on what aspects of a model's functionality, against what standards? What will be the standards for admissibility of statements or other evidence, or insights generated by AI and/or relied upon (or rejected) by humans? How will we assess its credibility or **authenticity**[9]? How will we define "properly" and "well-crafted" with respect to data collection and use case selection?
- **Explainability**: The feature of existing machine learning techniques that describes the degree to which the resulting model is able (or more likely unable) to provide an understandable explanation for how it reached its outputs. Generally, the more complex the models get, the harder it is for them to identify how and why they produce their particular outputs. In practice, this concept includes technology techniques that help approximate how a model produces an output as well as analog descriptions of processes for how decisions are made or influenced by AI model outputs. One aspect of some machine learning techniques is that the resulting model is unable to provide a human-understandable reason for its decision or recommendation. How does the ability to understand or interrogate an AI model impact the evidentiary value of its output? What is an expert in this area? What levels of uncertainty in model outputs will be acceptable to the courts, and under what circumstances? Conversely, how do different level(s) or ranges of model certainty relate to different standards of proof? (i.e., when does X, Y, or Z degree of correlation [perhaps minus X%, Y% or Z% of uncertainty] equal some legal level of proof (e.g., "clear and convincing," "preponderance of the evidence" or even "beyond a reasonable doubt))" Will this be a matter left up to states, courts, or judges to decide for each unique case? Or will standards be crafted and enforced? Should these standards be strict or flexible?
- **Law:** The substantive law will start to change and reflect the introduction of AI into broader society and commerce. Questions will arise about how to apply existing legal principles to new fact patterns involving AI (e.g., jurisdiction, employment discrimination, insurance claims, contract disputes, intellectual property claims, professional malpractice claims, negligence, tort liability, consent, duties of care, *mens rea*). Issues related to the right to cross-examination, and/or claims of proprietary protections are already emerging. Moreover, questions will arise that require entirely new legal principles. There are already efforts in some jurisdictions to regulate AI and the data that fuels it, as well as efforts to apply existing standards in new contexts (See Annex A).

[9] That which is genuinely who or what it is represented to be.

- **Liability and Standards of Proof:** Moreover, courts will soon be dealing with issues related to the degree to which humans can or should rely on technology. Should humans be supervising or reviewing certain outputs and decisions recommended by AI systems before they are acted upon? Who is responsible for malfunctions or **hacking**[10] of technologies using AI? Expect arguments about the inability of parties to entirely understand or control certain AI-powered functions and litigation about what is reasonable.
- **Remedies**: How traditional remedies map onto harms relating to AI is an enormous open question. Redressing breaches of discrete contract or license terms may prove relatively straightforward but dealing with complex inputs and supply chain questions and ongoing conduct, and injunctive relief more generally, will not be straightforward. Moreover, issues related to the degree to which humans are, or should be supervising or reviewing outputs[11] and/or who is responsible for malfunctions or hacking of technologies using AI will arise. Expect arguments about the inability of parties to entirely understand or control certain AI-powered functions.
- **Litigants:** Already, litigants have new analytical tools at their disposal to assess their claims and the courts themselves. Likewise, tools are available to gain insights into clients, counterparties, potential jurors, adversaries and witnesses.

3. Elements and Types of AI Systems

Each element of an AI system, including the human and non-technical elements, contributes to and affects AI performance across the entire **AI Lifecycle**. The AI Lifecycle describes the full process and steps by which an AI model or system is commissioned, designed, developed, trained, implemented, monitored and decommissioned. There are many ways to describe the AI Lifecycle, but it is common to look at the following phases:

1. **Use case selection:** choosing and articulating the business or other applied purpose or goal for which AI tools will be used and defining the context for their application.
2. **Design/Selection**: designing the AI system and the model types that are fit for purpose and appropriate to the use case and the context in which the tools will be used.
3. **Design/Development**: engineering the AI model, selecting and preparing appropriate data sets, training[12] and testing the model.

[10] The act of gaining access to a system without authorization, often with the intention to manipulate or destroy it. Hacking can be used as a means of testing a system's security and functionality.

[11] We Are the Loop, Not Just In It: Success Needs a Focus on Humans At Every Step in the AI Lifecycle

[12] Model training is an essential stage in AI model development that involves continuously feeding training data to the model and reviewing and/or adjusting certain model features or variables. Most models need many rounds of training and tuning.

4. **Deployment and Monitoring**: putting the AI system to work in the real world, training the users and field engineers, and watching how (and as needed, adjusting) the AI system functions relative to its purpose and any future proposed purposes.

It is relatively simple to recite these phases, but actual AI systems are engineered under a wide variety of conditions, with overlapping stages and levels of capability. The foregoing is simply an ideal, to set the stage.

3.1. Use Case Selection

Use cases are contextual problems that AI systems are meant to address. Ideally, using an AI-powered tool will increase accuracy, efficiency and cost savings, perhaps reduce human bias, create new work opportunities, and so forth. AI is not (yet) helpful to address every sort of problem, and not every sort of problem requires AI to solve it.

- If an AI tool has been selected for use, someone, somewhere, should be able to explain why it has been selected (even if it is intended as an experimental solution) and what it is intended to execute, optimize, or solve for.
- One should likewise be able to articulate the grounds or measures against which the tool's functioning and definition of success will be evaluated. For example, is the tool being assessed in relation to optimal human performance, or some other higher standard?

Interestingly, the same sort of AI tool might work very differently in one setting than another, so understanding the purpose for which AI is deployed and the context can be very instructive to understand later impacts and assess levels of process and care taken in the development and deployment (or redeployment) of AI systems.

This is also a good step at which buyers can apply **procurement standards for AI tools** to assess the demonstrated capabilities (beyond any hype by sellers) and fitness for purpose, available resources to implement the tool internally, appropriate data, controls and so forth. To the extent sellers are under any obligation to make disclosures, limit uses or support buyers, the procurement process is one moment where these issues will arise and **risk management**[13] and mitigation efforts can begin.

[13] Establishing policies and practices to help manage and mitigate the risks posed by an AI system.

3.2. Design/Selection

Examples of AI Applications. Assuming an appropriate use case has been selected, several AI techniques and types of models have been demonstrated to work very well. These include (but are in no way limited to):

- **Natural Language Processing (NLP)** is a machine learning technique that analyzes large quantities of human text or voice data (transcribed or acoustic) for specified features, including meaning, content, intention, attitude, context and many others.
- **Sentiment Analytics**, for example, uses NLP and other AI techniques and inputs to correlate features of language (or facial movements, gaze, etc.) or other **Biometric Data** (heart rate, retinal patterns, etc.) to analyze or predict humans' affective or emotional states.
- **Image Recognition** uses AI models (usually computer vision) to detect, recognize and understand (living or material) objects in pictures or video. **Facial Recognition** tools, for instance, use computer vision technology to recognize (and in some cases, identify) a human face via biometric measurements and characteristics. Some facial recognition technologies simply 'match' known faces with faces already in a database and others also correlate those matches to other information to predict the identity of a person.
- **Predictive Analytics** is the overarching category of statistical tools and models that can use and analyze historical data to make predictions about the future to inform decision-making, such as machine learning (ML) systems. These predictions can be benign (which movie to recommend tonight) or more consequential (which loan application to recommend approving) or deeply inappropriate in a judicial setting (which defendant is most likely to engage in a particular behavior).
- **Bots** and **Chatbots** use NLP to understand and automate verbal or written responses to customer questions and needs, simulating an interaction with a human customer service representative. Many customer service bots also integrate sentiment analytics, for instance, to manage call routing.
- **Immersive or augmented experiences**: **Augmented reality** is the overlaying and/or integration of computer-generated visual, auditory, haptic, or somatosensory information and phenomena onto real-world objects and interactions. **Virtual reality** experiences occur in entirely computer-simulated worlds, often immersive. It is through AR and VR technologies (currently most often through a headset or goggles) that one can access the **metaverse**, that while lacking a widely agreed upon definition, generally refers to a robust virtual 3-D world in which individuals can simulate real world (or novel) activities. The metaverse is just beginning to become a mainstream environment and expectations are that it will grow in acceptance and sophistication very quickly, and because it will represent complete 'worlds' it will raise every sort of legal issue.
- **Connected Devices or Internet of Things/Sensors** are networks of devices, such as wearables (e.g., smart watches), thermostats and sensors on a refrigerator, industrial

production line, or in a car that feed data to AI models in order to improve performance or generate predictions about the devices or the users of the devices.

- This technique is applicable in many contexts, including **Remote Patient Monitoring, Smart Cities** and **Autonomous Driving** (the latter of which of course also includes image recognition and several other types of AI).
- **Synthetic Content** and **Deep Fakes** are the byproducts of AI-enabled technology that renders hyper-realistic content that can appear so real that it can exceed the capacity of human beings to detect it as fake. **Generative AI** uses a model's own underlying logic and training to generate new artificial outputs or datasets. This is great in gaming or rebalancing datasets, but less great in spreading believable misinformation. Questions around authentication and authenticity, provenance and reasonable reliance are inevitable (to name a few).
- **Robotic Process Automation (RPA)** is the process of combining AI tools (especially for spatial navigation and task ordering) with very advanced machine engineering to enable more advanced functionality. Three common administrative use cases for RPA include filling out forms, synthesizing data and organizing electronic files.

Most AI tools combine several analytic or mechanical techniques to develop an approach to a problem set. For instance, a smart phone can use facial recognition technology, in combination with security features and other phone features to unlock the device. Or a robot can do a backflip or pick and pack at a warehouse by combining AI tools for spatial navigation and task ordering with very advanced machine engineering. A résumé screening tool could use NLP to select certain features in candidate information, assess the sentiment of the résumé's author and perhaps predict the duration of a candidate's future tenure, prospects for advancement, or any number of other criteria.

Framing, Goal Alignment and Metrics Selection. What an AI model selects for and how it optimizes on a goal will be a function either of what goals the humans have set for it, or some combination of features and goals that the model itself has determined are significant, based upon its training and function. Accordingly, design issues stem from the AI model itself and the system surrounding it, including the involvement of humans in the development of the model and as consumers of the model's outputs.

AI models are very sensitive to how instructions and objectives are chosen, described, framed and articulated and how well a model's goals are aligned with well-considered human intentions. **Small differences in how a question is asked can have a big impact on what predictions / outputs an AI model *produces*.** (For instance, "identify images with atypical cells" is likely to produce different output than "identify images with no atypical cells" or "identify images with atypical cells to a certainty level of X").

Likewise, data scientists can determine whether a model is designed to **interpolate**[14] data to produce prediction outputs, or rather to **extrapolate**[15] to generate its predictions. The distinction can be critical (the correlation vs. causation questions will be especially important to judges). A conscious (or unconscious) skew for **false positives**[16] or **false negatives**[17] can have a big impact, especially if predicting human behaviors. Further, whether to frame questions to solve for high levels of certainty will be more or less appropriate depending upon the use case. Regardless, how the question is asked should be the byproduct of forethought and close alignment between the entity seeking to use the model and those designing it, to ensure the questions are being framed properly and the results are appropriate and interpreted properly.

Level of Human Supervision. What sort of model to select or AI system to design should reflect how humans will interact with it (i.e., what some refer to as humans being "in" "on" or "out" of the "loop").[18] For instance, an AI model designed to optimize telecom network traffic presumably will not require as much human engagement as an AI model designed to optimize surgery schedules at a hospital or efficiently approve and deny loan applications.

Intended Human/Machine Outcomes. Depending upon the use case and capability of the AI system, the model itself will have different purposes:

- **Augmenting human or machine performance:** AI models that assist or supplement humans or existing technologies in executing a task, potentially by reducing the time and energy necessary to complete it. Most AI today falls into this category.
- **Replacing human performance:** AI model supplants the human performer. This remains a minority of applications today.
- **Replacing machine performance:** AI model outperforms existing technologies, resulting in their substitution.
- **Algorithmic decision-making** and **Automated decision-making** are terms that are frequently used but lack widely agreed upon definitions. In general, they refer to an AI model permitted or instructed to determine outcomes but can also less specifically refer to an AI system or model that facilitates, recommends or influences human decisions or outcomes.[19] In some situations, this can be appropriate and in other situations less so, especially where humans are substantially impacted by the outcomes and do not have the opportunity or ability to override or challenge the model or its decisions. Cautions have been raised about tools that rely on AI models to make decisions about loans or

[14] Interpolation refers to making discrete inferences within a constrained set of known data points.

[15] Extrapolation refers to using known values or AI model outputs to estimate the value of unknown variables.

[16] Indicates the presence of a fact or condition when that fact or condition is actually not present (a "type one error").

[17] Indicates the absence of a fact or condition when that fact or condition is actually present (a "type two error").

[18] We Are the Loop, Not Just In It: Success Needs a Focus on Humans At Every Step in the AI Lifecycle

[19] The distinction between recommending and decision-making is almost always context dependent.

benefits eligibility for instance, without some level of human oversight and/or rights to challenge adverse decisions.

3.3. Design/Development

Model design/development focuses on the inputs that are used to construct the model. Those inputs, broadly speaking, consist of data and algorithms (logic).

Data. Data are essential to AI models and are sometimes described as the fuel or oxygen that power AI. Data are powerful and increasingly plentiful, but data can be highly differentiated, and their value is highly contextual, so perhaps not best described as commodities. **Datasets** refer to collections of data used to train AI models. The datasets on which AI models are trained and then operate vary widely in terms of size, quality and source. Data that are useful to address some questions (your address to deliver your package) may be entirely inappropriate to address other questions (your address to determine your eligibility for college).

The success of an AI model depends upon the appropriateness and quality of its data. So how data are selected and processed is a critical inquiry. Datasets come in a few key forms:

- **Labeled** or **Unlabeled:** Labeled data is tagged or annotated (usually by humans but sometimes by AI itself) with meaning and categories that reflect its understood contents, characteristics and features. Ex: A photo of a cat with the corresponding annotation "cat." Unlabeled data lacks any information tags, annotations, or classifications. Ex: A photo of a tractor lacking any annotation.
- **Structured** or **Unstructured:** Structured data is data that is organized in standard formats and categorized contextually and relationally. Ex: Credit card transactions, financial transactions, forms, addresses, dates, some forms of legal discovery. Unstructured data is raw data or data points that lack any relational meaning or significance to one another. Ex: Photos, videos, images of handwritten notes and unrelated documents. Structured data contains **features**, which are measures or variables, often qualitative, that affect or distinguish categories of data points (e.g., height, weight, level of education) in structured data. The features in a dataset will depend both on the type of data used, the context of the data and the purpose or goals of the analysis. These decisions can impact how an AI model generates correlations between features and predictions on the basis of those correlations. For example, is someone's height, or the color of the roof of their home correlated with grade point average (GPA)? Or credit score? Hence, this feedback loop warrants inspection.
- **Derivative/Recursive:** Data that is itself produced, processed, altered or augmented by the AI model. Ex: Insights from a predictive analytical model that are fed back into a dataset of images that have been stylized to add or remove complexity.
- **Synthetic**: Artificial data that is generated to have the same characteristics as real data but do not tie back to real people or events.

- **Cleaned** or **Uncleaned:** Cleaned data has been labeled, relabeled, or reviewed to ensure that data points are properly organized, appropriate, or equivalent. Sometimes outlying or irrelevant data may be removed. This process of "cleaning" data can be labor and time intensive and will differ by dataset. Uncleaned data is a dataset in its raw or unfinished form.
- **Dirty:** Data can be inaccurate, incomplete, or contain errors or volumes of extraneous material.
- **Representative:** A dataset is representative if it accurately and appropriately reflects or measures the population or phenomena it is intended to capture, relative to the purpose for which it will be used. This can be a question, for example, where a model that will impact a broad community is trained on data that excludes elements of that community or does not represent the whole of that community. The specific standards for representativeness are highly contextual.
- **Open** or **Proprietary:** Open data is publicly available data that can be accessed or used with or without a license. Proprietary data belongs to and is reserved for its owner, who may decline to license or disclose it. This can be a question, for example, where an AI tool is developed and trained on proprietary data to which the eventual customer does not have access, complicating the customer's ability to audit or assess the tool's performance.
- **Healthy** or **Poisoned:** Healthy datasets are those containing accurate and complete data. Datasets can be vulnerable to tampering, manipulation and "**pixel hacking**."[20] Negligent use or overtly adversarial efforts to poison datasets,[21] even at very subtle levels, can profoundly affect the integrity of the AI model's outputs.
- **Merged:** Used to describe individual datasets that have been combined to form larger, more representative or more complete data sets. Merging data sets requires systematic matching and cleaning and this process of "**munging**" by data scientists can be time-consuming and require great care.

AI model design requires distinct datasets at three distinct stages:

- **Training Data** is the historical data used to develop and teach an AI model the logic and pattern recognition to generate desired predictions in the future. Ex: data sets of historical job performance or images of atypical skin cells.
- **Test Data** is the data used to evaluate how well a trained model is performing once it is built and before it is released. Ex: a new résumé or image of a skin cell.

[20] Pixel hacking refers to pixels embedded or manipulated, often unknowingly, in digital images and communications that can alter dramatically an AI model's performance and/or track and collect personal information often without informed consent.

[21] Poisoned datasets are those that have been tampered with, manipulated, or otherwise distorted in a manner that negatively impacts the quality and utility of the dataset.

- **Production Data** is the data used by the model once it is released for operational or commercial use. Ex. A particular job applicant's résumé or patient's biopsy.

Algorithms: AI model development requires selecting from among today's proven techniques, which practically speaking, are likely to be flavors of **machine learning (ML)**[22] or **deep learning**,[23] which are used to process enormous quantities of data to detect patterns in the data and then make predictions based upon those patterns. Both ML and deep learning can work on millions of variables at one time, vastly exceeding the processing capacity of humans.

What is an algorithm?

An **Algorithm** is mathematical logic (sometimes called "rules") used to organize, evaluate and assess matches and patterns in data. An algorithm is coupled with data to make an **AI Model.** The AI model is developed by and impacts human experience. **AI Systems** describe the AI models (which, as discussed, are composed of algorithms and data), along with the humans, organizations and other technologies that make up the whole ecosystem in which the AI model is running. It is important to broaden the lens to consider the *human and social systems* around the models, because how well those systems operate directly determines how well the models and the technology really work—and the impacts they really have—in applied settings.

How are AI models created?

AI Lifecycle

Use case selection: choosing and articulating the business or other applied purpose or goal for which AI tools will be used and defining the context for their application.

Design/Selection: designing the AI system and the model types that are fit for purpose and appropriate to the use case and context in which the tools will be used.

Design/Development: engineering the AI model, selecting, and preparing appropriate data sets, training and testing the model.[24]

Deployment and Monitoring: putting the AI system to work in the real world, training the users and field engineers and watching how (and as needed, adjusting) the AI system functions relative to its purpose and any future proposed purposes.

The key techniques for ML in use today include (but again, are not limited to):

- **Supervised Learning**: An AI technique in which humans (or pre-existing models) instruct a new model on the relationships that exist within and between certain pairs of input-output data. Ex: a model is trained to identify email spam in an inbox by correlating

[22] The most common subset of AI used to process enormous quantities of data to detect patterns and make predictions based upon those patterns.

[23] A subset of machine learning that uses pattern analysis to process millions of variables for prediction or sensing, often without supervised optimization.

[24] Model training is an essential stage in AI model development that involves continuously feeding training data to the model and reviewing and/or adjusting certain model features or variables. Most models need many rounds of training and tuning.

language that has previously been identified as 'spam' with language it finds in new emails.

- **Unsupervised Learning**: An AI technique that does not instruct the model about how data points are related (if at all), but rather relies on the model to derive those relationships and discover patterns in the data. Ex: a model is exposed to all purchasing data and discovers unobvious behavioral patterns and associations between consumers that predict the likelihood of a future purchase.
- **Reinforcement Learning**: The ability of an AI model to learn how to perform a series of discrete steps correctly and consistently within its use case environment because a user "rewards" or "punishes" the model depending on the accuracy of its actions. Ex: a model that targets advertisements receives positive reinforcement when users click on the ad, and negative reinforcement when they ignore or identify the ad as irrelevant.
- **Neural networks**: A set of algorithms modelled (roughly) to mimic the human brain that identify the relationships between data points in a network-like fashion. Neural networks[25] contain a multitude of interconnected artificial neurons, or nodes, that are assigned weights and biases and are organized into at least three layers (input, hidden and output). When a node in the hidden layer receives an input stimulus above a determined threshold, it activates and fires information to a node in the output layer. Some neural networks are so complex and accommodate so many variables that their outputs are difficult to interpret, interrogate and/or explain.

3.4. Deployment and Monitoring

AI models and systems are increasingly deployed in commercial, academic and government settings. Indeed, one would be hard-pressed to identify a part of the economy or area of human life that AI does not already touch. It therefore is critical to understand this technology's functions and impacts in applied settings.

Outputs. AI models produce **predictions** by correlating information and recognizing patterns from past events or instances (data) with new data to forecast the likelihood of an event or instance occurring in the future – meaning AI models offer probabilities and carry inherent uncertainty. Predictions can occur at a general or population level (e.g., demographic predictions to inform consumer product design), or they can be very specific as to one person

[25] Generative Adversarial Networks (GANs) are an unsupervised deep learning method that can generate hyper realistic content. GANs are used for unsupervised deep learning methods, including generating realistic photographs or examples for image datasets, performing text-to-image and image-to-text translations, aging faces, and creating emojis. GANs employ two neural networks – a generator that creates new examples and a discriminator that attempts to distinguish these fake, and often poor quality or unrealistic, images from the real image data fed into the AI system. Through this interaction, the generator learns to create more convincing and high-quality images, eventually tricking the discriminator into believing they are part of the real image data.

Convolutional Neural Networks (CNNs) are a machine learning technique often used for classifying or generating images and videos in unstructured or unlabeled data.

(hyper-personalized) or molecule (e.g., Ms. Smith is predicted to want to buy this brand of running shoes this afternoon). And everything in between.

Predictions fall into two categories:

- **Recommendations:** A suggested outcome or course of action informed by a model's predictions. Ex: what movies or shoes a particular person may enjoy, or job to which they should apply, based upon their attributes and historical behavior (as compared to others similarly situated, as determined by the model).
- **Insights:** AI system outputs containing actionable knowledge or information that can support a variety of human or technical functions, including decision-making. Ex: whether someone is likely to quit a job, or skip a workout or miss a doctor's appointment, based upon their attributes and historical behavior (as compared to others similarly situated, as determined by the model).

Obviously, how these recommendations and insights are used, who has access to them, whether the object of them has consented to them being generated, used, shared, or aggregated, all are important questions.

Performance and Continuous Monitoring. The following sections discuss how to assess the performance of AI models and their key limitations and risks. For purposes of the AI lifecycle however, the critical point is that AI models, and therefore systems, must be continuously monitored (see below). They are designed to improve, adjust and reform, so an innate characteristic of AI models is their outputs change over time and can drift from their original levels of performance, and even their original objectives.

As discussed above, the appropriateness of any of these outputs will depend greatly on the context for which the outputs are being used, how those outputs are being delivered and by whom. Many outputs become training and production data in other contexts, so assessing the compounding benefits or risks in a complete model can be difficult.

Human Accountability and Training. Humans are essential to how AI systems are deployed and monitored. The specifics for how humans do this, when and how extensively, will turn on the context of the use case:

- **Oversight and Accountability:** Generally, refers to how AI systems are governed and by whom, and how risk and responsibility are allocated across these processes.
- **Training**: Coursework, lectures or training for employees, customers and managers who need either basic or specialized training to understand, operate and manage frontier technologies.
- **Community Standards and Duties of Care**: How traditional duties of care for product development or handling will map onto AI systems, or the handling of their outputs.
- **Decommissioning:** The act of removing an AI model from deployment once its function is complete or performance is degraded, or to correct errors or mitigate risk.

- **Subsequent or re-use:** Determining whether an existing AI model is fit for purpose in a new context, for a new purpose or use case or otherwise expanding its capacities through **transfer learning**.[26]
- **Continuous Model Monitoring**: The need for consistent vigilance over and review of a model's activity to track changes and fitness for purpose. Many AI models are designed to improve their own performance over time, with more data and more experience. This is a hugely powerful and beneficial feature of AI, and it also creates new and different risks associated with models that are fit for release in their initial form but depart from their original purposes or performance over time.

4. Performance of AI Systems

AI system performance is measured both in terms of the intended system functioning and the effect on human and social arrangements. As discussed, AI systems are valued for their potential to generate insights at a speed and scale, and sometimes in a dimension, that exceeds human capacity. AI done well promises to reveal entirely new ways for humans to understand the world around (and within) them. It will produce insights that represent breathtaking advancements. As we will discuss in the next section, however, it also risks catastrophic outcomes if not properly scrutinized and managed.

An AI system should primarily be assessed on how well its model's predictions align with its articulated objectives and the trained outputs, and how appropriate the model is to the task, and how well tested and trained the model and people are, relative to its impact. **There is a growing understanding that AI systems are not perfect, and that perfection is not an appropriate standard; what *is* the appropriate standard will differ by (at least) application, jurisdiction and legal or regulatory regime.**

In general, AI systems' performance can be assessed on several technical (vs. impact) measures, the significance of which will shift depending upon the use case:

Efficacy	The continued ability to accurately execute a defined task
Accuracy	The ability to produce a correct or true value relative to a defined parameter.
Validity	No agreed definition, but generally refers to an AI model or system that produces repeatable outputs or measurements that are appropriate and within a range of design objectives.
Reliability	The ability of an AI model to produce consistent results over time.

[26] Re-using and re-applying a pre-trained model for a new problem set, use case, or in a context for which it was not initially designed or trained.

Cost	The ability to reduce or eliminate costs associated with human and/or machine labor.
Time	The ability to reduce the time it takes to complete a task.
Improvements	The ability to produce outcomes relative to the model's stated objective and/or today's baseline performance. How improvements are measured will be a function of the AI system's purpose, context and impact and how the AI system works relative to historical experience with humans or other technologies.
Velocity	The rate at which an AI model operates and can produce outputs (which factor into a model's impact).
Scale	The scope and size of the population (people or otherwise) impacted by the AI model (this also factors into a model's impact).
Precision	The accuracy of model or system outputs in meeting intended definitions or thresholds for relevance or significance.

5. Inherent Limitations and Risks of AI Systems

AI systems possess several inherent limitations, some of which are technical and some of which are social. For instance, while AI systems can recognize and establish **correlations** between data points, they cannot understand or establish **causality**, or whether variables are related by associations that would be meaningful to human understanding (e.g., identified the wolf from the dog based on its physical stance in a photo), or relationships between unimportant or inappropriate data points (e.g., identified the wolf from the dog based on the presence of snow in the photo).

- Interestingly, unexpected associations may be instructive or detrimental. This is one form of an **unintended consequence,** as would be the introduction of new social impact or security vulnerabilities. Many models, however, cannot explain how they generated their insights, and these models suffer from **opaqueness**[27] or a lack of **transparency**.[28] Sometimes outputs are **uninterpretable**[29] and always, models will include some margin for **uncertainty**, **error** and some form of **bias** (see below).

In considering the limitations and risks of AI, a few important observations:

[27] Used to describe a model that cannot necessarily explain to humans how it arrives at its insights or how humans should (or should not) rely on those insights.

[28] Concept associated with communicating how a model or system operates, generates outputs or is expected to make an impact, including known limits.

[29] Interpretability concerns whether a model output or explanation is understandable and intelligible to humans.

- This field is still in development and there is a lot still to learn.
- Very (maybe most) often, even developers do not or cannot know exactly how an AI model or system will perform when launched or over time.
- There is no single standard for what "good" or "responsible" or "reasonable" AI looks like or for that matter, what "bad" AI looks like.
- Even the best-designed, most well-intentioned AI will have some inherent limitations. And the same may be said of humans.

Broadly speaking, the limitations and risks of AI models fall into the following categories: **Uncertainty and Error, Bias, Opaqueness** and **Security** vulnerabilities. The causes and effects of these vulnerabilities overlap, influence and reinforce one another, so it is difficult to separate them neatly. With that caveat, we summarize each in turn.

5.1. Uncertainty and Error

Because AI models operate by producing predictions – even very good ones – their outputs contain a root, innate, irreducible element of uncertainty or error. So do humans of course. Because AI models are being developed to address ever-more sophisticated social/legal problems, there is an inexorable element of innovation and experimentation, inevitably increasing uncertainty and complexity.

There are no agreed upon or common legal standards for what levels of uncertainty or error are tolerable, or what amounts to negligence or recklessness. The answers here will turn on a variety of factors and influences: laws, court decisions, community guidelines and industry / corporate practice. In some areas, existing laws and regulations will apply, and in others, new legislation may be required. Global, federal, state and local standards will differ in substance and in the pace of adoption. Much of the work in AI governance is an effort to articulate and establish minimum standards in the meantime and to address the gaps in requirements. But how the law will be practically enforced, and according to what standards, will be a matter for the courts (See also: Annex A below).

Due to the velocity and scale at which AI systems operate, which greatly outpace the scope of human processing, even small uncertainty and error rates can result in substantial impacts. Accordingly, it is possible for AI models with a lower error rate than humans to nonetheless pose a greater risk.

Uncertainty and error in an AI system can arise at any point in the AI lifecycle, and be the result of technological or human factors, or the product of how humans and technology interact.

- **Technological uncertainty and error:** The idea that most AI models involve machine or deep learning techniques, which at their core calculate correlations between features and recognize patterns in datasets to produce predictions that are not absolute or entirely certain. Those predictions will have a certainty score (e.g., this image depicts a cat to a level of 98% certainty).

- **Human uncertainty and error:** AI systems are themselves vulnerable to ordinary human struggles to communicate clearly and avoid error-prone outcomes; likewise, while AI can reduce human errors in some areas, it will not eliminate human errors.
- **Human and technology interaction:** The more humans and AI interact, the more they will impact each other and begin to create their own, novel dynamics. Humans will over-rely on some AI and neglect to override or disregard its predictions; humans will also under-rely on AI and reject reasonable suggestions that could improve outcomes. Humans will instruct AI models to do brilliant things, and terrible things.

Critically, uncertainty and error can be introduced at the design and development stages, or appear long after an AI model is developed, launched and deemed fit for purpose. For instance, a model initially operating with great accuracy, reliability and validity in one particular context may over time begin to drift from its intended performance standards or objectives, or may assume false associations or replicate the undesirable behaviors of the human or machine actors in its use case environment. Some examples (non-exhaustive):

- **Inappropriate use case selection:** AI tools and techniques are not (or not yet) the solution for every problem or able to answer any question. These tools are well-suited for particular sorts of problems for which high-quality data are available; they are not appropriate for all questions. Likewise, a use case could be inappropriate because it contravenes law, ethics, norms or just conflicts with another business objective, or because a model that is designed for one use case is then deployed in another, for which the model is not suited.
- **Goal misalignment and framing problems**. AI models must be designed with an objective. How that objective (and any limits on that objective) is conceived and expressed will have enormous effects on how well the model performs or does not. Goal-setting as regards AI is subject to several vulnerabilities:
    - **Human:** Humans are notoriously imprecise communicators, and even when they are paying close attention and trying hard to align a model's goals with well-considered outcomes, individuals vary widely in their language, lived experience, culture, age, gender and circumstance, all of which can impact a model's results. For this reason, among others, AI development teams that are personally and professionally diverse can increase the number of perspectives and the development team's chances of anticipating how framing, language and other assumptions might work well or go awry. In some environments, an overly broad or acutely specific **description of instructions** will not matter very much, or might even be beneficial (e.g., discovering more instances of pre-cancerous cells); in others, however, the consequences might be more dire (e.g., failing to

detect cancerous cells based on a narrow set of **parameters**[30] and **design objectives**).[31]

- **Human-Technology:** Technologists ultimately translate human language instructions into a machine-readable format. That process requires coders to make many small decisions about how to describe something in and with code, and thus produces opportunities for some nuance to be lost, or perfectly natural assumptions, misunderstandings, or biases to creep in.
- **Technology:** AI models do not have context or limits unless they are designed and coded to have them. AI models cannot read between the lines or begin to infer human intention. Some scholars have therefore advocated for AI models that are designed to ask questions when something is unclear or in conflict. AI models are also susceptible to **function creep**, occurring when the AI system expands beyond the scope, purpose, or use case context for which it was designed or in which it was deployed, often, but not always, without proper validation for the new use.
- **Incomplete, inappropriate, or unrepresentative training or production data:** AI model training data can be incomplete or over- or under-representative of instances, objects, or populations. A dataset used for training or production should be properly constructed and scoped relative to a well-crafted use case and a well-framed objective. Misalignment in this area can lead to errors and bias. Likewise, datasets that are compromised or not properly cleaned can introduce errors and biases.

- **Overfitting and Underfitting:** An AI model that too narrowly or strictly **correlates** data features can miss important relationships and associations in data that may sit just outside those parameters (over-fitting). Conversely, an AI model that too loosely correlates data features can assign significance between features where none actually exists (under-fitting). Both are forms of system error that can impact the utility of an AI model.
- **Drift and Degraded Performance:** Drift can occur when an AI model diverges from its instructions and training, either by no longer performing its programmed tasks or performing unarticulated tasks and producing new outputs or insights. Some believe that models are inevitably prone to drift and therefore to degrading performance over time. This can occur when a model initially operating with great accuracy, reliability and validity in a particular context begins to overlearn common outcomes (and exclude important edge cases), assume false associations, or replicate the undesirable behaviors of the human or machine actors in its use case environment (e.g., the chatbot Tay

[30] The characteristics or rules of a model or AI capability that define its articulated use, performance and limits under specific circumstances.

[31] Design objectives refer to parameters and instructions that are outlined or defined in accordance with the purpose or goals for an AI system.

became racist and abusive within a matter of days, due to how humans interacted with it). For this reason, AI models require **continuous model monitoring**[32] and oversight.

- **System or impact error.** When an AI model is fed bad data or used in a novel or ill-defined situation for which it was not designed or trained, or for a purpose to which it cannot legitimately transfer its learning, this can lead to detectible and undetectable errors and unintended consequences.

5.2. Bias

All AI models – even very good ones – operate by processing volumes of historical data, organizing them by rules and using labels often provided by humans, and so contain some element of bias.[33] As do humans of course.[34]

Some bias is necessary to the proper functioning of any AI model, as that is what enables the model to distinguish between typical and atypical cells in an image of the skin or recommend shoes that fit one's style from those that do not. The social, legal and ethical questions arise in how AI models function when they predict matters with social implications, such as who should receive government benefits or be advanced in an employment context.

Bias can come into an AI system at any stage of the AI lifecycle and be a byproduct of how humans, data and the algorithms are functioning, and certainly how all three interact.

- **Human bias.** Whether implicit or explicit, existing human biases will make their way into AI systems through the sorts of use cases selected, how questions for AI models are framed, which data are selected, how model instructions are converted into code, and even how model outputs are interpreted or communicated to affected communities. The impact of human bias may appear as "data" and may impact individuals or communities of individuals.

Not all disparate impact is unlawful, but some is, and how to apply existing standards to algorithmic outputs is a matter for the courts.

- **Data bias:** Datasets contain historical information and reflect historical patterns that if themselves biased, will encode those biases in models that are trained on that data. Datasets compile information about historical events and as such are artifacts of human history and experience. Some data, even if biased, reflect neutral matters (e.g., how a particular communications network performs over time, or pictures of cats). Some data, however, reflect interpretive and socially reflective events (e.g., hiring practices over time, rates of criminal recidivism, or pictures of faces). Likewise, datasets can be under or over-representative of certain communities, such that the same model can work well

[32] Consistent vigilance over and review of a model's activity to track changes in the model and fitness for purpose.

[33] Bias refers to a preference or tendency for one thing over another, whether for reasons of rational choice or intentional or unintentional discrimination.

[34] NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, 2022.

for some populations and less well for others (e.g., facial recognition might have a lower uncertainty rate for white or male faces and higher uncertainty rate for brown or female faces if trained on datasets that do not adequately represent darker complexions or women).

- **Algorithmic bias:** The tendency of an AI model to produce outputs that perform differentially because of how an algorithm is built or functions with the data on which it is trained or operated. The algorithm itself can encode or learn bias. Usually this is a byproduct of earlier decisions by technologists, data selection, learning techniques or training models.

How to identify and mitigate unacceptable or unfair bias is a substantial challenge. Increasingly, methods and tools are rapidly advancing and being developed to aid in the assessment of AI models and their performance against certain defined parameters (e.g., is the model producing outputs that are "fair" as measured against some standard). These tools and techniques are evolving along with the models themselves and need to be evaluated independently for their applicability and quality.

5.3. Opaqueness

AI models – even very good ones – operate at a level of complexity and by means that are not always understood or understandable. **Explainable AI** is a distinct field of technical work dedicated to revealing model explainability and how certain AI models work and produce predictions.

- **Technological opacity.** The degree to which AI models or their results are or can be explained, or alternatively operate in a so-called **black box**,[35] will differ by the type of model, use case and context. Accordingly, what it is reasonable to expect by way of explanations or transparency will also differ. These difficulties can arise in assessing what are valid and accurate outputs, instructions and training. Difficulties interrogating a black box AI model's prediction-making process can inhibit the ability to appeal decisions based upon those predictions. Black box AI models also increase the possibility for errors, particularly those unrecognizable to or difficult to detect by humans, to go long undiscovered.
- **Human opacity.** The degree to which AI systems are explainable is a little more within the range of typical inquiries about design, inclusion, process, training, oversight and contextually appropriate (or required) levels of transparency and disclosure. While how AI model outputs are used, what appeal processes are available, and so forth, are more within the ambit of human management process, there will remain certain unknowable and unforeseeable outcomes with AI models and systems.

[35] Phrase used to describe a technological system that is inherently opaque, has inner workings or underlying logic not fully understood, or results and outputs that cannot be explained.

5.4. Security

The speed and scale of AI systems can introduce both new **security risks** as well as solutions. AI models' heavy reliance on massive datasets and conversely, humans' reliance on AI models also put new capabilities in the hands of both benign and malign actors. The full range of intersection between AI and cybersecurity is beyond the scope of this chapter. However, the presence or absence of adequate security safeguards or risk indicators can be relevant and impactful.

6. Who is Regulating AI Systems?

Both domestic and international government and non-government organizations have expressed interest in regulating AI systems. The AI regulatory landscape is vast, fragmented and changing very rapidly. (See Annex A).

7. Conclusion

Change is the only constant. This is a fast-moving area and from a judicial perspective it will be critical to understand the role of humans in creating and living with this change, what sorts of questions to ask about AI/ML, and how to keep up with developments and capabilities, and rapidly changing norms. All of which is to say, judges will determine how rights and disputes involving these technologies are adjudicated, and those decisions will impact every one of us in new ways, in all aspects of life.

Further complicating and magnifying the impacts of these changes is that an increasing number of people are splitting time between the physical world and virtual worlds.

As the number of people working, playing and living in the metaverse increases, so too will the number of complaints and legal disputes. Complaints of sexual assault[36] and verbal harassment and concerns around tracking and misuse of personal data, money laundering and cyberbullying are on the rise.[37] Judges, therefore, will not only be tasked with remediating conflicts in the natural world but with applying laws to a world that is still largely being developed.

[36] The metaverse has a groping problem already

[37] Crimefighting in the metaverse

Annex A: Global Regulation Review[38]

Both domestic and international government and non-government organizations have expressed interest in regulating AI systems. The AI regulatory landscape is vast, fragmented and changing very rapidly.

What follows is illustrative and anyone needing to understand the state-of-the-art will need to research and update this list regularly.

A.1. United States

The United States is taking a sector-specific approach to regulating AI systems so far, as opposed to introducing omnibus legislation.

- **Federal Executive.** A November 2020 Memo released by the Office of Management and Budget (OMB) proposed ten principles for U.S. Government agencies to consider when developing their regulatory or non-regulatory plans for the commercial development and use of AI. The OMB noted that agencies should only regulate when necessary, affirming the Government's interest in extending and tailoring existing legislation to AI, rather than introducing new rules. Several **U.S. Government Agencies** have started to undertake this work, including, but not limited to, the Government Accountability Office (GAO), the Federal Trade Commission (FTC), the Department of Defense (DoD) and its Defense Innovation Unit (DIU), the U.S. Food and Drug Administration (FDA), the Department of Veterans Affairs (VA), the National Geospatial-Intelligence Agency (NGA), the Department of Health and Human Services (HHS) and the Department of Homeland Security (DHS). In September 2021, a Commissioner of the Equal Employment Opportunity Commission (EEOC) expressed intentions to examine AI bias and discrimination in employment and hiring. In May 2022, this office released a guidance document titled, "The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees."
- In October 2021, the Office of Science and Technology Policy (OSTP) announced plans to develop an AI Bill of Rights, or a set of principles and standards for the use of technologies. In December 2021, the National Institute for Science and Technology (NIST) announced plans to develop an AI Risk Management Framework (RMF) that will outline voluntary standards for managing and reducing AI risks. NIST released its second draft of the RMF and accompanying RMF playbook in August 2022.
- **Federal Legislative.** In 2020, Congress passed The Artificial Intelligence Initiative Act (AI-IA), mandating the OSTP, NIST, the National Science Foundation and The Department of Energy to carry out certain activities related to AI. Activities include investing in AI research, creating AI advisory committees and helping develop measurements and standards that support the innovation of commercial AI applications. In support of these

[38] As of September 1, 2022

goals, the National AI Advisory Committee and Subcommittee on AI and Law Enforcement were created in Spring 2022.

- **States.** Several states, including California, Illinois, New Jersey and Washington have introduced and/or passed significant legislation aimed at regulating AI systems. Many bills deal with particular AI applications, such as facial recognition.
- **Municipalities.** Roughly two dozen U.S. municipalities have banned government use of facial recognition software, citing concerns around accuracy and discrimination. In late 2021, the New York City Council became the first to enact a law requiring organizations to notify candidates and employees when an automated AI decision tool is used to assess their eligibility for hire or promotion. This law comes into effect in January 2023.
- **Civil Society.** US civil society groups are also very active in their calls for AI regulation. An October 2021 letter written to Ambassador Susan Rice by the Leadership Conference on Civil and Human Rights cautioned that "without top leadership in the White House focused on ensuring that civil rights and civil liberties issues are front and center in the development of technology and AI policies, the development of these policies will continue to center the voices of the tech industry and place civil rights at the margins." The 100-plus civil society group signatories, including the ACLU, NAACP, LDF, Common Cause, urged regulation of AI and other frontier technologies.

A.2. Global

- The **Organization for Economic Co-operation and Development (OECD)** has identified over 700 global AI policy initiatives from 60 countries, territories and the EU. The four most common policy instruments include national strategies, agendas and plans, emerging AI-related regulation, public consultations of stakeholders or experts and networking and collaborative platforms. In May 2019, the OECD adopted *Principles on Artificial Intelligence*, which although voluntary, have been readily adopted by OECD member states and others. The principles aim to promote AI that is innovative and trustworthy and respectful of human rights and democratic values.
- The **European Union's** proposed Artificial Intelligence Act (AIA) seeks to categorize AI models according to three risk categories: high, limited and minimal. Documentation, disclosure and transparency requirements correspond with these risk levels, ranging from voluntary to stringent and obligatory. A fourth category, unacceptable risk, prohibits the deployment of certain high-consequence and high-risk AI applications, such as social scoring systems. An open statement signed by over 150 European civil society groups calls on officials to amend the Artificial Intelligence Act (AIA) to recognize and "addresses the structural, societal, political and economic impacts of the use of AI" and protect democratic rights and values.

A.3. Specific Countries

- **Australia** adopted a National AI Ethics Framework and Principles in November 2020 to support the responsible development of AI. Australia's AI Action Plan (June 2021) has budgetary support for investment in Cooperative Research Centre projects, PhD scholarships, and other initiatives to increase the supply of AI talent in Australia.
- **Brazil** published its AI Strategy in April 2021. In July 2021, its House of Representatives proposed an AI bill to create a legal framework for the development and use of AI by the government, private sector, civil society, NGOs and individuals. The framework requires alignment with human rights and democratic values (including reference to Brazil's data protection law). It also requires impact studies and reports.
- **Canada** was the first country to introduce direct regulation of government use of AI. Canada adopted a Directive on use of Automated Decision-Making by Federal Government, effective April 2020. It sets limits, requirements and standards for the Canadian federal government's use of automated decision-making tools, including for transparency and procedural fairness. The Directive applies to how the federal government may use AI and sets new obligations on companies licensing or selling AI systems with these capabilities to the federal government.
- **China** has grown very active in its AI governance and regulation efforts through several national organizations. The Cyberspace Administration of China focuses on developing standards and rules for online algorithms and soliciting public opinion. In 2021, the CAC published the most formal articulation of the country's AI rules, the Internet Information Service Algorithmic Recommendation Management Provisions, which is effective March 2022. The CAC has likewise released a set of Guiding Opinions on Strengthening Overall Governance of Internet Information Service Algorithms. The China Academy of Information and Communications Technology is responsible for developing tools to test and certify "trustworthy AI" systems. 2021 publications include a White Paper on Trustworthy AI and a Trustworthy Facial Recognition Evaluation Program. The Ministry of Science and Technology assembles AI ethics review boards within organizations and research institutes to help establish AI ethics principles. 2021 publications include Guiding Opinions on Strengthening the Ethical Governance of Science and Technology and Ethical Norms for New Generation Artificial Intelligence.
- **Germany** created a new Study Commission on "Artificial Intelligence Social Responsibility and Economic, Social and Ecological Potential." This Study Commission will investigate how AI and algorithmic decision-making will affect society. It is made up of 19 Members of Parliament and 19 AI experts. We expect the Committee to produce a report with policy recommendations.
- **India** is studying AI but has yet to take meaningful action. In November 2020, NTI Aayog, a government public policy think tank, recommended creating a consortium of ethics councils focused on developing sector-specific guidelines on privacy, security and ethics and a National AI Marketplace. In February 2021, the think tank released a draft

document in consultation with World Economic Forum's Centre for the Fourth Industrial Revolution to encourage the responsible use of AI in India. India's Personal Data Protection Bill (2019), which may be implemented in 2022, seeks to protect and secure citizens' personal data and encourage growth in the country's digital economy.

- **Japan** was the first country to raise as a matter of official policy the need to create governance standards for the development and implementation of AI. During its G7 Presidency, Japan called for international guidelines and principles to govern the AI R&D in 2016. The Japanese Government held a consultation earlier in 2021 on AI Governance Principles for organizations.
- **Singapore** is likewise taking a sector-specific approach to AI regulation. Principles released by the Monetary Authority of Singapore (MAS) in 2018 aim to promote fairness, ethics, accountability and transparency (FEAT) in the use of AI and data analytics in Singapore's financial sector. Since 2021, the MAS has partnered with industry (Veritas Initiative) to pilot these principles, resulting most recently in the February 2022 release of five white papers:

    1. FEAT Checklist;
    2. Fairness Assessment Methodology;
    3. Ethics and Accountability Assessment Methodology;
    4. Transparency Assessment Methodology; and
    5. Assessment Case Studies.

    In support of the Government's AI work and expanding the commercial development and deployment of AI, Singapore created an industry-led Advisory Council on Ethical Use of AI and Data in 2018. Singapore's proposed AI governance framework, released in 2019, was complemented by the 2019 publication of its Trusted Data Sharing Framework. These two documents, focused on ethical and Responsible AI, became operational in 2020 when the Government partnered with the World Economic Forum to create an Implementation and Self-Assessment Guide for organizations.

    In May 2022, Singapore released the world's first AI Governance Testing Framework and Toolkit in partnership with the World Economic Forum.

- The **United Kingdom** is taking an approach like that of the US, focusing on AI investment and sector-specific standards. This perspective was affirmed in a 2018 House of Lords report, which recognized that sector-specific regulators are best equipped to evaluate AI regulation and that blanket AI regulation is inappropriate at this time. In 2019, the UK became the first country to pilot AI procurement guidelines, which focused on increasing the adoption and use of Responsible AI in the public sector. In 2021, the Government released a National AI Strategy outlining how investments in Responsible AI development and AI skill and job creation will promote broader economic growth. As

part of this national strategy, in January 2022, the Government announced plans to pilot a new initiative focused on shaping global technical standards for AI.

A.4. Global Civil Society

- In November 2021, the United Nations Educational, Scientific, and Cultural Organization (**UNESCO**) General Conference adopted the UNESCO Recommendation on the Ethics of Artificial Intelligence. This international framework for the development and use of AI is aligned with the UN's Sustainable Development Goals and is grounded in human rights. In January 2022, UNESCO and the Innovation for Policy Foundation (i4Policy) held a workshop to help validate their ten building blocks for inclusive AI policies that were developed through five multistakeholder and cross-disciplinary workshops:

    1. Raise awareness and demystify AI
    2. Define AI and be clear on terminology
    3. Use an expert group to determine the AI landscape
    4. Provide clarity on the participatory AI policy process
    5. Co-create and consult AI policy
    6. Make sure participation is followed by influence on outcomes
    7. Make AI policy agile and flexible
    8. Protect Human Rights, embrace Data Protection and Ethics Guidelines
    9. Combine the AI Strategy with an Action Plan
    10. Monitor and evaluate throughout the policy cycle

- **The World Economic Forum** is a leader in Responsible AI and AI governance efforts. In support of its mission to ensure AI systems promote privacy, accountability, equality and inclusion, the Forum actively encourages collaboration between the public and private sectors, regularly publishes guidelines, toolkits (e.g., for Boards, C-Suite, Procurement and Human Resources), and self-assessments for organizations, and assembles teams of subject-matter experts to confront novel technology challenges. Its Centre for the Fourth Industrial Revolution brings together teams from 15 countries to answer a central question: "How can we maximize the benefits of technology but avoid all potential risks?" The Forum's Global AI Action Alliance (GAIA) likewise works to increase global awareness and adoption of trusted, transparent and inclusive AI systems.
- In 2021, The **World Health Organization** (WHO) issued its first global report on artificial intelligence. The report offers six principles that the WHO believes will ensure AI supports the international community's public interest:

    1. Protecting human autonomy;
    2. Promoting human well-being and safety and the public interest;

3. Ensuring transparency, explainability and intelligibility.
4. Fostering responsibility and accountability;
5. Ensuring inclusiveness and equity; and
6. Promoting AI that is responsive and sustainable.

Glossary

Acronyms	Term	Definition	Synonyms & Related Terms
	Acoustical Processing	A sub-discipline in physics that measures sound, vibration, ultra- and infra-sound in various kinds of matter (solid, liquid and gas).	
	Affect Analysis	Analysis on language, gait, posture, facial movements or other characteristics to analyze an individuals' affective or emotional state.	Sentiment Analysis
	AI Agent	An AI Model (or set of AI models) in production that operates to generate "independent" recommendations or decisions.	AI Model
	AI Lifecycle	The full process and steps by which an AI model or system is commissioned, designed, developed, trained, implemented, monitored and decommissioned.	
	AI Model	An AI model is the product of applying an algorithm (or set of algorithms) to data in order to optimize on a particular goal and/or produce insights about a particular question or objective.	AI; Algorithm
	AI System	The ecosystem that includes AI models (themselves composed of algorithms and data), along with the humans, their organizations and any other technologies associated with their AI Lifecycle.	AI
	Algorithm	Mathematical logic (sometimes called "rules") used to organize,	Logic, Math, Computation,

Acronyms	Term	Definition	Synonyms & Related Terms
		evaluate and assess matches and patterns in data.	Program, Code, Mathematical Model
	Algorithmic Bias	The tendency of an AI model to produce outputs that perform differentially because of how an algorithm is built or functions with the data on which it is trained or operated.	Bias; Data Bias (See also paper on "Legal Research and Judicial Analytics")
ADM	**Algorithmic Decision Making / Automated Decision Making**	No widely agreed upon definition. Generally refers to an AI model permitted or instructed to determine outcomes, but can also less specifically refer to an AI system or model that facilitates, recommends or influences human decisions or outcomes. The distinction between recommending and decision-making is almost always context-dependent.	
GAI/General AI	**Artificial General Intelligence**	AI capable of autonomous existence across a full range of capabilities, with the ability to establish its own objectives, adjust to circumstances, learn and follow through on complex instructions.	
AI	**Artificial Intelligence**	No widely agreed upon definition. AI is both a concept and a category of technology tools that are powered by advanced mathematical models and data that can augment, replicate or improve upon the type of human cognitive task that otherwise	Machine Learning, Augmented Intelligence, Intelligent Automation

Acronyms	Term	Definition	Synonyms & Related Terms
		requires thinking, beyond calculating.	
ANI/Narrow AI	**Artificial Narrow Intelligence**	Narrow AI refers to models aimed at performing discrete tasks and functions.	Machine Learning, Augmented Intelligence, Intelligent Automation
ASI	**Artificial Super Intelligence**	AI models that are fully autonomous and self-executing in ways that surpass human intelligence and behavioral ability; they may or may not be "conscious" and may or may not be regenerating.	
AR	**Augmented Reality**	The overlaying and/or integration of computer-generated visual, auditory, haptic, or somatosensory information and phenomena onto real-world objects and interactions.	
	Augmented Human or Machine Performance	AI models that assist or supplement humans or existing technologies in executing a task, potentially by reducing the time and energy necessary to complete it. Most AI today falls into this category.	
	Authenticity	That which is genuinely who or what it is represented to be.	

Acronyms	Term	Definition	Synonyms & Related Terms
	Bias	Preference or tendency for one thing over another, whether for reasons of rational choice or intentional or unintentional discrimination.	Algorithmic bias, Data bias
	Big Data	The enormous, and exponentially growing, amount of complex data (structured, unstructured, and semi-structured) that is driving AI models today.	
	Biometrics	Data points captured and measured from human bodily functions and characteristics, including heart rate, retinal patterns, facial geometry, gait, speech etc.	
	Black Box	Phrase used to describe a technological system that is inherently opaque, has inner workings or underlying logic not fully understood, or results and outputs that cannot be explained.	
	Bot	An AI agent that mimics human behavior, capacity or processes.	
	Chatbot	An AI agent that simulates human customer service or conversational interactions by automating verbal and written responses to communicate with humans.	
	Cleaned	Data that has been labeled, relabeled or reviewed to ensure that data points are properly organized, appropriate, or	

Acronyms	Term	Definition	Synonyms & Related Terms
		equivalent. Sometimes outlying or irrelevant data may be removed.	
IOT	**Connected Devices/ Internet of Things/ Sensors**	Networks of devices, such as wearables (e.g., smart watches), thermostats and sensors on a refrigerator, industrial production line, or in a car that feed data to AI models in order to improve performance or generate predictions about the devices or the users of the devices.	
	Continuous Active Learning	A subset of machine learning in which models continuously learn and adapt based upon increasing amounts of iterative feedback.	
	Continuous Model Monitoring	Consistent vigilance over and review of a model's activity to track changes in the model and fitness for purpose.	
CNN	**Convolutional Neural Network**	A machine learning technique often used for classifying or generating images and videos in unstructured or unlabeled data.	Deep learning; neural networks
	Data Bias	Datasets are compiled of data that is collected from information about historical events and as such are artifacts of human history. Data bias refers to datasets that can encode biases if used to train models that then predict future eligibility based upon those historical patterns.	Algorithmic bias. See also the paper, *Artificial Intelligence, Legal Research, and Judicial Analytics.*

Acronyms	Term	Definition	Synonyms & Related Terms
	Datification	Pervasive collection, generation, storage and analysis of data that commodify data and data analysis in the form of unique predictions and insights.	
	Dataset	Collections of data used to train AI models.	
	Data Protection	Efforts and obligations to secure, obtain and use data according to prescribed standards.	Security; Privacy
	Decommissioning	The act of removing an AI model from deployment once its function is complete or performance is degraded, or to correct errors or mitigate risk.	
	Deep Learning	A subset of machine learning that uses pattern analysis to process millions of variables for prediction or sensing, often without supervised optimization.	Machine learning, CNN, GAN, Neural network
	Derivative/Recursive	Data that is itself produced, processed, altered or augmented by the AI model.	
	Design Objectives	Parameters and instructions that are outlined or defined in accordance with the purpose or goals for an AI system.	Instructions as Code / Description of Instructions
	Dirty Data	Describes data that can be inaccurate, incomplete, or contain errors or volumes of extraneous material.	

Acronyms	Term	Definition	Synonyms & Related Terms
	Drift	Drift can occur when an AI model diverges from its instructions and training, either by no longer performing its programmed tasks or performing unarticulated tasks and producing new outputs or insights.	
	Error	**Type One:** Indicates the presence of a fact or condition when that fact or condition is actually not present. **Type Two:** Indicates the absence of a fact or condition when that fact or condition is actually present.	See also: Annex C of *Artificial Intelligence, Trustworthiness, and Litigation.*
	Explainability	The feature of existing machine learning techniques that describes the degree to which the resulting model is able (or more likely unable) to provide an understandable explanatiaon for how it reached its outputs. Generally, the more complex the models get, the harder it is for them to identify how and why they produce their particular outputs. In practice, this concept includes technology techniques that help approximate how a model produces an output as well as analog descriptions of processes for how decisions are made or influenced by AI model outputs.	

Acronyms	Term	Definition	Synonyms & Related Terms
XAI	**Explainable AI**	A distinct field of technical work dedicated to revealing model explainability and how certain AI models work and produce predictions.	
XR	**Extended Reality**	Refers to all categories of environments generated or enhanced by AI.	
	Extrapolation	Using known values or AI model outputs to estimate the value of unknown variables.	
	F_1	The Harmonic Mean of Recall and Precision, often used in Information Retrieval studies to measure the effectiveness of a search or review effort, which accounts for the tradeoff between Recall and Precision.	See “Precision” below.
FRT	**Facial Recognition**	The use of computer vision technology to recognize (and in some cases, identify) a human face via biometric measurements and characteristics.	
	False Negative	Indicates the absence of a fact or condition when that fact or condition is actually present (a "type two error").	
	False Positive	Indicates the presence of a fact or condition when that fact or condition is actually not present (a "type one error").	

Acronyms	Term	Definition	Synonyms & Related Terms
	Features	Measures or variables, often qualitative, that affect or distinguish categories of data points (e.g., height, weight, level of education) in structured data.	
	Function Creep	When an AI system expands beyond the scope, purpose, or use case context for which it was designed or in which it was deployed, often, but not always, without its proper validation for the new use.	Scope Creep
	Gait Analysis	A form of biometric data analysis that evaluates the ambulatory motions of a person or mobile being.	
	Gaze Analysis	A form of biometric data analysis that evaluates retinal and eye movement to analyze patterns in motion.	
GANs	**General Adversarial Networks**	An unsupervised deep learning method that can generate hyperrealistic content. (GANs) are used for unsupervised deep learning methods, including generating realistic photographs or examples for image datasets, performing text-to-image and image-to-text translations, aging faces and creating emojis. GANs employ two neural networks – a generator that creates new examples and a discriminator that attempts to distinguish these fake, and often poor quality or	Deep learning, neural networks

Acronyms	Term	Definition	Synonyms & Related Terms
		unrealistic, images from the real image data fed into the AI system. Through this interaction, the generator learns to create more convincing and high-quality images, eventually tricking the discriminator into believing they are part of the real image data.	
	Generative AI	Type of AI that uses a model's own underlying logic and training to generate new artificial outputs or datasets.	
	Hacking	The act of gaining access to a system without authorization, often with the intention to manipulate or destroy it. Hacking can also be used as a means of testing a system's security and functionality.	
	Healthy Data	Datasets that contain accurate and complete data.	
	Human Bias	Whether implicit or explicit, existing human biases can make their way into AI systems through the sorts of use cases selected, how questions for AI models are framed, which data are selected, how model instructions are converted to code, and even how model outputs are interpreted or communicated to affected communities.	

Acronyms	Term	Definition	Synonyms & Related Terms
	Human Uncertainty and Error	AI systems are themselves vulnerable to ordinary human struggles to communicate clearly and avoid error-prone outcomes; likewise, while AI can reduce human errors in some areas, it will not eliminate human errors.	
	Interpolation	Making discrete inferences within a constrained set of known data points.	
	Interpretability	Describes a model output or explanation that is understandable and intelligible to humans.	
	Image Recognition and Analysis	The use of AI models (usually computer vision) to detect, recognize and understand (living or material) objects in pictures or video.	
	Insights	AI system outputs containing actionable knowledge or information that can support a variety of human or technical functions, including decision making.	Outputs
	Labeled	Used to describe data that is tagged or annotated (usually by humans but sometimes by AI itself) with meaning and categories that reflect its understood contents, characteristics and features.	
ML	**Machine Learning**	The most common subset of AI used to process enormous quantities of data to detect	See also: "Technology Assisted

Acronyms	Term	Definition	Synonyms & Related Terms
		patterns and make predictions based upon those patterns.	Review." Compare: AI
	Merged	Used to describe individual datasets that have been combined to form larger, more representative or more complete data sets. Merging data sets requires systematic matching and cleaning and this process of "munging" by data scientists can be time-consuming and require great care.	
MVS	**Metaverse**	No widely agreed upon definition, but generally refers to a robust virtual 3-D world in which individuals can simulate real world (or novel) activities.	
MR	**Mixed Reality**	A hybrid digital and physical setting that combines computer-generated visual, auditory, haptic, or somatosensory information and phenomena with real-world objects and interactions.	Augmented reality
	Model Alignment	The process of fitting a model to optimize goals and insights about a particular question or objective framed and described by humans.	
	Model Training	An essential stage in AI model development that involves continuously feeding training data to the model and reviewing and/or adjusting certain model features or variables. Most models need	

Acronyms	Term	Definition	Synonyms & Related Terms
		many rounds of training and tuning.	
NLP	**Natural Language Processing**	A machine learning technique that analyzes large quantities of human text or voice data (transcribed or acoustic) for specified features, including, but not limited to, meaning, content, intention, attitude and context.	See also: Sentiment analytics (below)
	Neural Network	A set of algorithms modelled (roughly) to mimic the human brain that identify the relationships between data points in a network-map like fashion. Neural networks contain a multitude of interconnected artificial neurons, or nodes, that are assigned weights and biases and are organized into at least three layers (input, hidden, and output).	Deep learning, CNN, GAN
	Opaqueness/Opacity	Used to describe a model that cannot necessarily explain to humans how it arrives at its insights or how humans should (or should not) rely on those insights.	
	Open Data	Publicly available data that can be accessed or used with or without a license.	

Acronyms	Term	Definition	Synonyms & Related Terms
	Output	The result or value produced by an AI model. Outputs are generated from combining input data with the model, and fundamentally are **predictions.**	Insights
	Overfitting	Used to characterize an AI model that too narrowly or strictly correlates data features can miss important relationships and associations in data that may sit just outside those parameters.	See also: Underfitting (below)
	Oversight and Accountability	Generally refers to how AI systems are governed and by whom, and how risk and responsibility are allocated across these processes.	
	Parameters	The characteristics or rules of a model or AI capability that define its articulated use, performance and limits under specific circumstances.	
	Pixel Hacking	Pixels embedded or manipulated, often unknowingly, in digital images and communications that can alter dramatically an AI model's performance and/or track and collect personal information often without informed consent.	
	Poisoned	Datasets that have been tampered with, manipulated, or otherwise distorted in a manner that negatively impacts the quality and utility of the dataset.	

Acronyms	Term	Definition	Synonyms & Related Terms
	Precision	The accuracy of model or system outputs in meeting intended definitions or thresholds for relevance or significance.	See also: Precision-Recall Trade-off (and F1) discussion in the paper, *Artificial Intelligence, Legal Research, and Judicial Analytics.*
	Predictions	AI model outputs that result from correlating information and recognizing patterns from past events or instances (data) with new data to forecast the likelihood of an event or instance occurring in the future – meaning AI models offer probabilities and carry inherent uncertainty.	
	Predictive Analytics	The overarching category of statistical tools and models that can use and analyze historical data to make predictions about the future to inform decision-making.	
	Privacy	No widely agreed upon definition, but generally the broad category of personal interests associated with being free from unauthorized observation, surveillance, or intrusion.	
	Probabilities	Calculations that predict the likelihood of the occurrence of a certain event.	

Acronyms	Term	Definition	Synonyms & Related Terms
	Production Data	The data used by the model once it is released for operational or commercial use.	
	Proprietary Data	Proprietary data belongs to and is reserved for its owner, who may decline to license or disclose it.	
	Recall	The amount or portion of data/documents deemed "relevant" from a review/search of such material/data.	See also: Precision-Recall Trade-off (and F1) discussion in the paper, *Artificial Intelligence, Legal Research, and Judicial Analytics.*
	Recommendation	A suggested outcome or course of action informed by a model's predictions.	Output
	Reinforcement Learning	The ability of an AI model to learn how to perform a series of discrete steps correctly and consistently within its use case environment because a user "rewards" or "punishes" the model depending on the accuracy of its actions.	
	Reliability	The ability of an AI model to produce consistent results over time.	
	Representative	A dataset is representative if it accurately and appropriately reflects or measures the population or phenomena it is	

Acronyms	Term	Definition	Synonyms & Related Terms
		intended to capture, relative to the purpose for which it will be used.	
RAI	**Responsible AI / Ethical AI / Trustworthy AI**	Concepts that are used to describe AI models and systems that are designed, built and operated with forethought and certain key attributes that protect human beings (such as fairness, safety, transparency and a respect for human autonomy) and denote a level of quality in their performance overall.	
RM or RMF	**Risk Management**	Establishing policies and practices to help manage and mitigate the risks posed by an AI system.	
RPA	**Robotic Process Automation**	Combines AI tools (especially for spatial navigation and task ordering) with very advanced machine engineering to enable more advanced functionality.	
	Sentiment Analytics	Use of **NLP** and other AI techniques and inputs to correlate features of language (or facial movements, gaze, etc.) or other biometric data to analyze or predict humans' affective or emotional states.	See also NLP (above)
	Stochastic	Refers to AI models whose performance and outputs include some inherent level of uncertainty or randomness.	

Acronyms	Term	Definition	Synonyms & Related Terms
	Structured Data	Data that is organized in standard formats and categorized contextually and relationally.	
	Supervised Learning	An AI technique in which humans (or pre-existing models) instruct a new model on the relationships that exist within and between certain pairs of input-output data.	
	Synthetic Content/Deep Fakes	An AI-enabled technology that renders hyper-realistic content that can appear so real that it can exceed the capacity of human beings to detect it as fake.	
	Synthetic Data	Artificial data that is generated to have the same characteristics as real data but do not tie back to real people or events.	
	Technological Uncertainty and Error	The idea that most AI models involve machine or deep learning techniques, which at their core calculate correlations between features and recognize patterns in datasets to produce predictions which are not absolute or entirely certain.	
TAR	**Technology Assisted Review**	In the context of information science, mechanical systems for finding pertinent data in large datasets. Also/specifically, the use of algorithms to define and then narrow searches (of large quantities of data).	See also: ML (above), the paper, *Artificial Intelligence, Legal Research, and Judicial Analytics.*

Acronyms	Term	Definition	Synonyms & Related Terms
	Test Data	The data used to evaluate how well a trained model is performing once it is built and before it is released.	
	Traceability	Refers to the availability of information related to an AI model's production and deployment, including, but not limited to, when and by whom it was created, the datasets on which it was trained, the population(s) and context(s) in which it is and was deployed, and its performance.	
	Training (Human)	Coursework, lectures or training for employees, customers, and managers who need either basic or specialized training to understand, operate and manage frontier technologies	
	Training Data	The historical data used to develop and teach an AI model the logic and pattern recognition to generate desired predictions in the future	
	Transfer Learning	Re-using and re-applying a pre-trained model for a new problem set, use case, or in a context for which it was not initially designed or trained.	Subsequent or re-use
	Transparency	Concept associated with communicating how a model or system operates, generates	

Acronyms	Term	Definition	Synonyms & Related Terms
		outputs or is expected to make an impact, including known limits.	
	Uncleaned	A dataset in its raw or unfinished form.	Unstructured data
	Underfitting	Used to characterize an AI model that too loosely correlates data features and can assign significance between features where none actually exists.	See also: Overfitting (above)
	Unlabeled	Data that lacks any information tags, annotations, or classifications.	
	Unstructured Data	Raw data or data points that lack any relational meaning or significance to one another.	
	Unsupervised Learning	An AI technique that does not instruct the model about how data points are related (if at all), but rather relies on the model to derive those relationships and discover patterns in the data.	
	Use Case Selection	Choosing and articulating the business or other applied purpose or goal for which AI tools will be used and defining the context for their application.	
	Validity	No agreed definition, but generally refers to an AI model or system that produces repeatable outputs or measurements that are appropriate and within a range of design objectives.	

Acronyms	Term	Definition	Synonyms & Related Terms
VR	**Virtual Reality**	Experiences that occur in entirely computer-simulated worlds, often immersive	