AUGUST 2021

AI Primer

The New York City Artificial Intelligence Primer

Mayor's Office of the Chief Technology Officer

nyc.gov/cto

First published

August 18th, 2021

Version

File version 1.3.0818-16

Type

Typeset in *Libre Baskerville* and *Public Sans*, with accents of *Space Mono*.

Online

http://on.nyc.gov/ai-primer

Accessible HTML version:

https://www1.nyc.gov/assets/cto/#/publication/ai-primer

Note

The NYC AI Primer is subject to applicable laws, rules, and regulations, including City procurement rules and processes. The City reserves all rights, including rights to postpone, cancel, or amend the AI Primer at any time. The City shall not be liable for any costs incurred in connection with the AI Primer.

Table of Contents

Introduction

Artificial intelligence (AI) is changing the human experience today, driving sweeping social, economic, and technological transformation that affects us all. The City of New York believes that an approach grounded in digital rights is necessary to maximize its benefits, minimize its harms, and ensure its responsible application. Moreover, establishing a clear understanding of what AI is, how it works, and what some of the key practical and ethical considerations are around its use is foundational to building a healthy AI ecosystem for New York City.

"We are drowning in information and starving for knowledge."

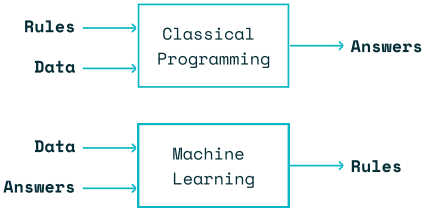
Rutherford D. Rogers
Former Chief of Research Libraries
The New York Public Library

One of the chief difficulties in the discourse on AI broadly is that claims — both positive and negative — are often exaggerated to the point of being misleading, and "AI" is also often used more as a marketing term than a precise description of the techniques used. Even among those working in the field, there can be inconsistency or disagreement with regard to scope, definitions, and priorities. To facilitate better policy, recognize both opportunities and risks, and evaluate claims made by others, New York City decision-makers require greater clarity on what "AI" means, what components can make up a system, the wide range of ways considerations like performance and accuracy, fairness, accountability, privacy, and security can come into play, and the complexity of weighing these factors against each other in any given situation.

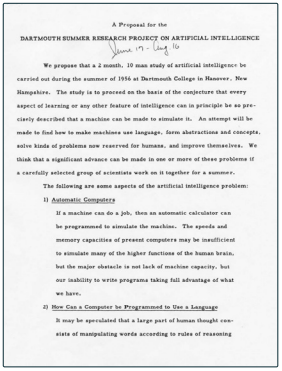
This document aims to help provide this foundation, primarily for an audience of technical, policy, or other decision-makers who are in or interact with New York City government. Importantly, this is a rapidly evolving field, and this should not be taken to be a comprehensive or final account. Ongoing engagement will be required to ensure local stakeholders are keeping pace with the technology, its use, and its consideration across society as each of these aspects continues to develop.

Computer chess was long one of the most visible testing grounds for AI technology. Shown here, World Chess Champion Garry Kasparov plays IBM's Deep Blue in 1996. *Photo: Laurence Kesterson*

AI is an umbrella term encompassing a range of technologies both sophisticated and simple that are used to, among other things, make predictions, inferences, recommendations, or decisions with data. The term "artificial intelligence" was first coined in the 1950s to describe efforts by computer scientists to produce general human intelligence and behavior in computers; these early efforts to create AI systems were largely "rule-based" to attempt to simulate human reasoning.



A simplified diagram comparing traditional software to machine learning, showing ML as a kind of new software programming paradigm. From F. Chollet, *Deep Learning with Python*, O'Reilly, 2017.



A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

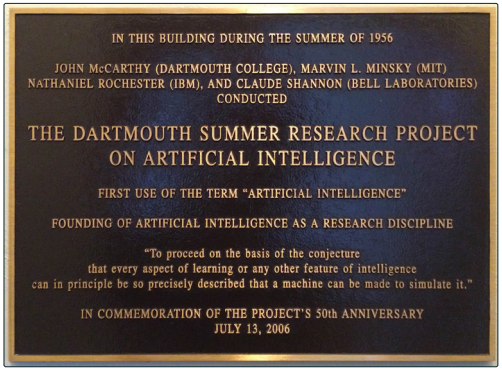
The following are some aspects of the artificial intelligence problem:

1) Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2) How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning



IN THIS BUILDING DURING THE SUMMER OF 1956

JOHN McCARTHY (DARTMOUTH COLLEGE), MARVIN L. MINSKY (MIT)
NATHANIEL ROCHESTER (IBM), AND CLAUDE SHANNON (BELL LABORATORIES)
CONDUCTED

THE DARTMOUTH SUMMER RESEARCH PROJECT
ON ARTIFICIAL INTELLIGENCE

FIRST USE OF THE TERM "ARTIFICIAL INTELLIGENCE"

FOUNDING OF ARTIFICIAL INTELLIGENCE AS A RESEARCH DISCIPLINE

"To proceed on the basis of the conjecture
that every aspect of learning or any other feature of intelligence
can in principle be so precisely described that a machine can be made to simulate it."

IN COMMEMORATION OF THE PROJECT'S 50th ANNIVERSARY
JULY 13, 2006

Initial proposal for the *"Dartmouth Summer Research Project on Artificial Intelligence,"* generally considered to be the founding of AI as a research discipline; a plaque commemorating the 50th anniversary of the event.

In traditional (non-AI) software, developers tell a computer exactly how to carry out a given task using precise, fixed instructions. This sequence of instructions is called an algorithm.[1] This approach works well for tasks like sorting a list of names or typesetting a book, but does not work well for problems like differentiating between photos of dogs and cats, reading the handwritten address on an envelope, or identifying fraudulent credit card transactions. Intuitively, there is far too much variation in these cases to handle with explicit rules, even though some of these tasks are easy for humans.

1 See, for example: T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms,* third edition, MIT Press, 2009; T. Roughgarden, Algorithms Illuminated, Soundlikeyourself Publishing, 2020; D. E. Knuth, *The Art of Computer Programming*, 2011, details at https://www-cs-faculty.stanford.edu/~knuth/taocp.html.

Among practitioners and specialists, "AI" now largely refers to the use of an approach called machine learning (ML), a way to write "software by example" by providing the computer with illustrative examples to "learn" from.[2] Machine learning uses data together with certain mathematical techniques to create computer programs. These techniques largely come from the fields of statistics, probability, mathematical optimization, and computer science, though increasingly tools from economics, the social sciences, and other areas in applied mathematics are used as well.[3] The computer is given a description of the task to be performed; data in the form of examples of what the correct results look like; a mathematical way of formalizing or expressing assumptions about how the data relates to the task called a "model"; and a learning algorithm indicating how to improve at the task by trial-and-error. The result is a "trained model" which can take new inputs for which the correct output or result is not known and guess the correct output. This guessed output is called a "prediction," which in this context is a technical term and often does not refer to predicting the future. The question of what kinds of outputs can in practice be effectively "predicted" with ML is itself a subtle topic and the subject of ongoing research.[4]

Because references to statistics or statistical language are so pervasive in AI, it is worth briefly addressing a potential point of confusion. In the context of government, the word "statistics" often refers to what are sometimes called "administrative statistics," or government's definition of relevant measures (such as poverty, employment, or race) and subsequent measurement to facilitate governance.[5] These include anything from measurements (or "statistics") related to people to those related to agricultural production.[6] While not unrelated, this is different from the kind of "inferential statistics" or "statistical inference" used in AI.

The process of building an AI system is described in more detail with concrete, practical examples below, along with associated ethical and policy considerations that arise.

2 This document focuses on a particular form of machine learning called "supervised learning"; there are also other areas in machine learning, including "unsupervised learning" and "reinforcement learning."
For additional general reading, see the *Further References* section at the end of this document.

3 Because machine learning draws on so many other fields, there are often multiple pieces of jargon referring to the same concepts, depending on the academic training of the person speaking or even the venue in which an academic publication appears. In addition, many pieces of technical jargon from statistics and machine learning also have distinct colloquial uses that can cause confusion, including "discrimination," "bias," "prediction," and more.

4 J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, *"Prediction policy problems,"* American Economic Review, 2015; S. Athey, *"Beyond prediction: Using big data for policy problems,"* Science, 2017; A. Narayanan and M. Salganik, *"Limits to Prediction,"* 2020, available at https://msalganik.github.io/cos597E-soc555_f2020/ .

5 A. Desrosières, *The Politics of Large Numbers: A History of Statistical Reasoning*, translated by Camille Naish, Harvard University Press, 1993.

6 The US government has a decentralized *"Federal Statistical System,"* spanning 125 agencies engaged, to some degree, in collecting data and producing such descriptive statistics, with 13 agencies whose primary mission is statistical work; the best-known include the *Bureau of Economic Analysis, Bureau of Labor Statistics, and the Census Bureau.*

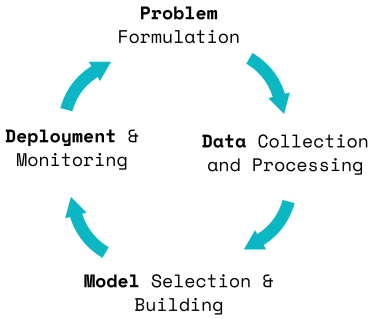
The AI lifecycle

The terms AI lifecycle or ML lifecycle describe the steps used to create a new machine learning system.

This generally involves the steps of identifying and precisely formulating a problem to be addressed; collecting and processing data; building a model; and deploying and monitoring the system.[7] In some cases, parts of this process are repeated iteratively: the team may collect new data and retrain or restructure the model every few months, or the system may automatically incorporate new data (learn) on an ongoing basis.[8] Because of the myriad challenges throughout this process, there are now entire companies that focus on offering products or services to aid other organizations even in single components of this lifecycle, from data labeling to system monitoring.

7 See, e.g., *Full Stack Deep Learning*, available at https://fullstackdeeplearning.com; *Stanford ML Systems Seminar Series*, available at https://mlsys.stanford.edu.

8 Systems that keep adapting (retraining) automatically over time are called "online" systems. There are pros and cons to this approach that need to be evaluated in context.



A simplified depiction of the lifecycle of an AI application.

This section outlines the AI lifecycle and discusses considerations that can arise at each stage. It will use mortgage lending[9] as a running example. Today, mortgage lenders often use ML algorithms to help make decisions about whether or not to approve a given loan application. This example was selected partly because it is both real and impactful, but also because it is simple while still exhibiting all of the different complexities discussed below.

The person or team going through the process below will be referred to as the "developer."

9 See, e.g., S. Trilling, *"Fair Algorithmic Housing Loans,"* Aspen Tech Policy Hub, 2020, available at https://www.aspentechpolicyhub.org/project/fair-algorithmic-housing-loans/. There are additional references on algorithmic lending and mortgage lending in citations through this document.

Problem formulation

Developers must first precisely formulate the problem.

This includes defining what the inputs and outputs are intended to be. In the case of mortgage loan decisions, the input would include a list of characteristics[10] of the loan application, potentially including information about the borrower, the financial terms of the loan, and details about the property, while the output could be taken to be "yes" (approved) or "no" (declined). This very common formulation is called "binary classification" because the input data is being classified into one of two classes or categories.[11]

There are other ways to formulate the problem. For example, the output could be a numeric[12] "risk score" from 0-100 that gives an estimate of the probability of loan repayment, and the developer would need to decide what role humans are intended to play in the process. It could be that the yes/no outputs are merely suggestions to help advise a human decision-maker; that the outputs are fully automated decisions; or that the system is designed to have three possible outputs, including a "maybe" option that prompts human review. Decisions like these are both necessary and subjective, and can have both practical and ethical implications.

It is essential to define what a model performing "well" means for the organization.

Often, this should be measured relative to some specified baseline (possibly the performance of a human team performing the same task), as systems can be flawed but still improve on the status quo enough that they are worth using. In mortgage lending, the lender may measure success based on corporate financial metrics (e.g., more loans get made with fewer borrowers defaulting on their loans); the system may also seek to behave similarly to humans but be faster or more transparent, or to improve on measures of equity and fairness. For example, researchers have found that although ML loan models do discriminate, they also can be up to 40% less discriminatory than face-to-face lending.[13]

"Some problems are better evaded than solved."

C. A. R. Hoare

10 This input list of characteristics can be visualized as a row in a spreadsheet, with one row per loan application and the columns corresponding to the different characteristics, such as borrower age and borrower income.

11 Binary classification is the formulation used for a very wide range of real systems, such as those that classify credit card transactions as valid or fraudulent or those that classify emails into valid or spam.

12 In addition to the output being one of a fixed collection of options (called classification) or a number (called regression), the outputs can also be much more complex structures. For example, in language translation systems, the input is a sentence in one language and the output is the correct translation in a different language; in a face detection system, the input is an image or video and the output may be the size and location of boxes that contain the faces in the image.

13 R. Bartlett, A. Morse, R. Stanton, and N. Wallace, *"Consumer-lending discrimination in the FinTech era,"* Journal of Financial Economics, 2021.

Beyond the technical aspects above, developers must consider the broader context in which the system will be deployed and if the concept motivating the system even makes sense.

It is important to understand that when systems perform poorly or have negative impacts in the real world, that can be caused by failures of problem formulation and conception rather than necessarily being a primarily technical issue.[14]

There are many points to consider at this stage, including the goals of the project, who it is intended to benefit, the stakeholders that should be consulted, auxiliary systems or processes that must be built around the system itself, and proactive consideration of possible malfunctions and their impacts on people. These are complex, interlinked, and context-specific, and there is no recipe for navigating them, so it is helpful to include a range of perspectives.

Depending on the application, engaging the public or other stakeholders may be a beneficial or necessary way to ensure this step is done effectively.

Community and public engagement, as well as participatory approaches, are covered in a subsequent section, but if it is appropriate to use engagement or participatory methods, this must be done in a thorough and careful way. For example, the developer must be willing to potentially reframe or even cancel the project altogether as a result of the input received, and it may be helpful to use quantitative or other specialized methods in order to effectively elicit and incorporate informed input. In short, there are a range of tools that can be brought to bear on this process that can complement or supplement familiar approaches like feedback forms and town halls.

14 See, e.g., V. Eubanks, "Automating Inequality: How high-tech tools profile, police, and punish the poor," St. Martin's Press, 2018; D. Kolkman, *'"F**k the algorithm"?: What the world can learn from the UK's A-level grading fiasco,'* LSE Impact Blog, London School of Economics, 2020, available at https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/.

Data

Although discussions of AI and ML often emphasize models or algorithms, it is well-known among practitioners that issues around data are often much more determinative of the success or failure of a project and can take up the vast majority of the time, effort, and cost.

One step is "data collection": For mortgage lending, a historical set of loan applications, each associated with a "ground truth" output value (such as approval status), must be obtained. This ground truth output may be available, may need to be inferred, or may require a new project in human annotation to produce — a process called "data labeling." Newer lenders would not have access to enough historical loan data to use ML at all, while older lenders may have paper records that need to be processed into a "machine readable" database, possibly requiring significant manual human effort. An important consideration is that for some tasks, the notion of "ground truth" may itself be somewhat or entirely subjective; this can both complicate the process of producing training data as well as result in downstream effects on the overall system built around it that may or may not be intended.[15]

Another step is "data cleaning," which generally refers to detecting and removing errors or inconsistencies in data. For example, loans may have been recorded with an inconsistent mix of 5-digit and 9-digit ZIP codes, some loan applications may include the borrower's gender while others don't, and even valid phone numbers can be written in many different formats. The properties may be represented by different brokers who provide data in inconsistent formats or based on different policies. Some data may be the result of forms completed by hand and later entered into a system, a common process which often introduces at least some errors. In addition, combining multiple datasets can be time consuming, error-prone, or even prohibitively difficult without standardized identifiers, such as social security numbers, license plate numbers, or Universal Product Codes.[16]

"The world is not the sum of all the things that are in it. It is the infinitely complex network of connections among them. As in the meanings of words, things take on meaning only in relationship to each other."

Paul Auster

15 For example, see A. Jeffries and L. Yin, *"To Gmail, Most Black Lives Matter Emails Are 'Promotions'"* The Markup, 2020, available at https://themarkup.org/google-the-giant/2020/07/02/to-gmail-black-lives-matter-emails-are-promotions.

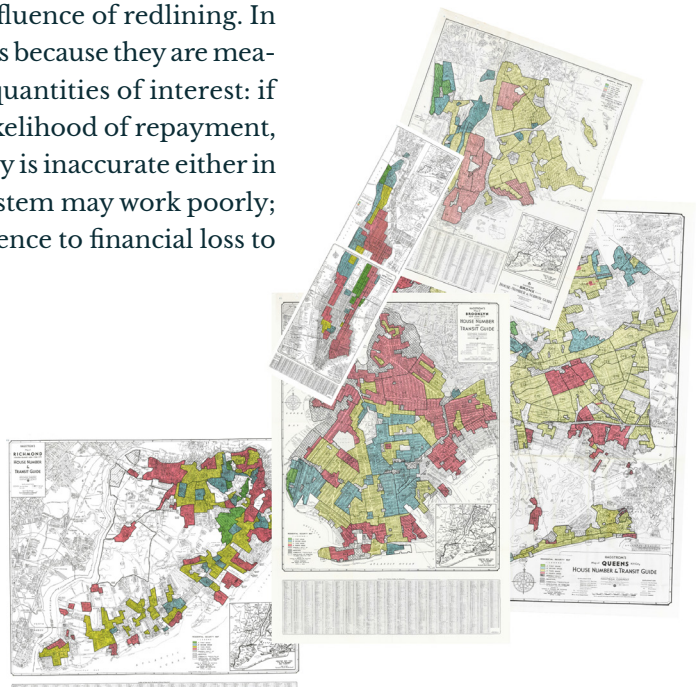
16 See https://www.gs1us.org/upcs-barcodes-prefixes/get-a-barcode/why-gs1-us.

These seemingly mundane tasks can be enormously time consuming and expensive and often require data engineering or domain expertise different from that needed to build models. For instance, some data may have been captured digitally at the source, possibly using an online form, but the specifics of how that form was designed — such as how a question was worded or how the input was validated — often shapes the actual data collected in ways that are opaque to a person looking at the data without that context. Lack of appropriate data for a task can lead to many problems, such as poor system performance or accuracy, data security breaches, or unfairness to particular groups, often for subtle reasons. For example, if there are very few people with certain attributes in the data, the outputs may be much less accurate for those groups because there is not enough data to go on.[17]

Certain values in the inputs or the output often serve as "proxies," which must be carefully considered to avoid undesired behavior.

This can be implicit or unintended: ZIP code can serve as a proxy for race, partly because of the historical influence of redlining. In other cases, developers explicitly use proxies because they are measurable and seem close enough to actual quantities of interest: if lenders are interested in borrowers' true likelihood of repayment, they may use credit scores instead. If a proxy is inaccurate either in general or in the context of the task, the system may work poorly; this can result in anything from inconvenience to financial loss to more serious harm.

17 For example, although facial recognition systems have been found to perform poorly for dark-skinned people in general, and dark-skinned women in particular, there have been subsequent studies showing that simply training the systems with datasets that are more representative appear to give significantly better performance on these affected groups. See, e.g., C. Romine (Director of Information Technology Laboratory at NIST), *"Testimony in Hearing on Facial Recognition Technology (Part III): Ensuring Commercial Transparency & Accuracy,"* Committee on Homeland Security, US House of Representatives, 2020; P. Grother, M. Ngan, and K. Hanaoka, *"NISTIR 8280: Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects"*, NIST, 2019; R. Puri, *"Mitigating bias in AI models,"* IBM Research Blog, 2018. For initial work on the discrimination itself, see J. Buolamwini and T. Gebru, *"Gender shades: Intersectional accuracy disparities in commercial gender classification,"* Conference on Fairness, Accountability, and Transparency, 2018.



Redlining typically referred to racial discrimination against particular neighborhoods when providing services or benefits; mortgage lending is one such example with a long history. This was often done explicitly, with "bad" neighborhoods outlined in red. See, e.g., R. Rothstein, *The Color of Law: A Forgotten History of How Our Government Segregated America*, Liveright, 2017.

Shown here, redlining maps of NYC from 1938.

Models

Models codify the developer's assumptions about the data and task at hand.

"All models are wrong, but some are useful."

George Box
Former President
American Statistical Association

The developer must next choose a model to use. Essentially, a model codifies a set of assumptions about the potential relationship between the inputs and the output, and can be thought of as a mathematical formula that combines the inputs with a set of adjustable numbers to produce a predicted output. These adjustable numbers are called "parameters" or "weights," to reflect that they indicate how much emphasis to put on different inputs for the given task. The types of models that are appropriate in a given situation are partly dictated by the problem formulation and the nature of the inputs and outputs; for example, one would use one type of model if predicting a yes/no binary output for mortgage loans and a different type if predicting a risk score or other value. Though some models can be very complex and have billions of parameters, extremely simple models (like linear regression, which dates back to the early 1800s) from traditional statistics can be very effective and are often used in practice as well.

Training a model involves using the data and a learning algorithm to set the parameters associated with the model. At the beginning of the training process, these are set to random values, which will of course result in the model initially producing poor predictions that do not match the labeled outputs; these discrepancies are incorporated in a learning algorithm to iteratively tune the parameters, resulting in a trained model. This model can take any input data in the same form as the training data, such as a new loan application with all the same attributes encoded as in the training data, and produce a predicted output (loan decision).

The key goal in machine learning is generally to make good predictions on new data that have not been manually labeled or previously seen.

Unlike what is often the case in social science projects, the goal is typically not to better understand existing historical data for its own sake using statistical methods. The main goal is to be able to do "well" at deciding what to do with new loans that have not yet been approved or denied, and one must define what "well" means.

There are many ways to measure and define performance.

For instance, in binary classification, there are four possible outcomes of a prediction; in mortgage lending, these correspond to correctly approving a loan ("true positive"), correctly denying a loan ("true negative"), incorrectly approving a loan ("false positive"), and incorrectly denying a loan ("false negative"). In the simplest case, one might care about predictive accuracy, which would just count the percentage of correct predictions. However, false positives and negatives are different types of errors with potentially different risks or impacts, so they may need to be accounted for differently. False negatives could result in both revenue loss for the lender and greater difficulty buying property for the borrower; false positives could lead to borrowers being given loans they cannot repay. When there are multiple criteria, there are often tradeoffs between them, and the developer needs to determine how to measure overall performance in line with organizational goals and potential broader impacts. To take the extreme case, one can easily drive either false positives or negatives to zero simply by denying or approving all loans, respectively, but this would obviously lead to the system performing very poorly on the whole.

In practice, developers will try out several different models on the same data and then choose one after seeing how well the models actually do on their specific problem with the data available. This decision can be based on several factors. In some cases, one may simply select the model with the best performance on the validation set (defined below). In other cases, developers may choose a different model that performs well enough but also has

other benefits, such as being easier to understand, inspect, audit, implement, or maintain. Depending on the setting, even the most sophisticated, best-performing model may perform too poorly to use in practice; conversely, as noted above, even a model that performs relatively poorly may still significantly improve on the status quo (possibly traditional software systems or human teams previously used to perform the task) either in performance or other ways, and still be well worth using. In short, model performance should be evaluated relative to an appropriately chosen baseline or status quo, not in a vacuum.

Performance is measured with respect to some particular choice of "validation" data.

To actually determine which model does best, there must be a way to evaluate how well it performs on data it was not trained on. Roughly, this is done by setting some of the labeled training data aside as "validation data"; validation data are not used in training and are only used to measure the performance of the model.

The following is critical: When a performance metric is reported for a machine learning model, that number should be understood as being evaluated on a particular set of validation data. If this validation data is not chosen appropriately, those metrics can give a misleading picture of how the model will perform in the real world. Because of this, and the varying ways in which performance can be measured, claims from vendors or in media reports that a system is "99% accurate" are often incomplete or outright meaningless without further details and explanation.

Deployment and monitoring

Unlike traditional software like a web browser, the performance and behavior of machine learning models often changes over time, so it is critical to carefully monitor systems once they are deployed.

"Do not be too positive about things. You may be in error."
C. F. Lawlor

These changes over time can happen for many reasons. Among the major reasons are that the data on which the system is used is, or becomes, different from the training data used to build the model, potentially because the underlying phenomenon being modeled itself changes.[18] For example, there may be inflows or outflows of certain types of people in an area; changes to housing supply or housing laws; or interest rates may rise to the degree that a different set of people are seeking mortgages. If a New York company buys a mortgage loan evaluation product from an outside vendor, it may be that that model is based on data from Florida, where loan data may look very different.[19] It is a virtual certainty that there will be many such shifts in the aftermath of COVID-19 in a range of domains, not only housing.

In other cases, actions taken based on the system's recommendations may be fed back into the system as new training data. This "online" re-training process is a means to keep a system up-to-date over time, but this process can also create the risk of feedback loops that could cause a model to, for example, increasingly only approve loans by existing homeowners, even if many first-time buyers are also likely to repay their loans.[20]

Finally, because of the complexity of engineering ML systems in general, there are a broad range of other practical considerations that arise in testing, monitoring, and maintaining these systems that must be considered.[21] As one example, there should not be any differences between the way data is processed to train the model and the way new data is processed before it is run through the model in production; any mismatches here can cause a range of potentially serious performance problems that can be difficult to detect and debug. Though this may seem obvious, this type of bug or mistake is very common in practice. As another example, it is important to be able to safely roll back to a previous version of a model that is known to work correctly.

18 These are sometimes referred to as "drift" or "shift," depending on the details.

19 This may sound far-fetched but is not hypothetical. In the context of medicine, see S. Lynch, *"The Geographical Bias in Medical AI Tools,"* Stanford Center for Human-Centered AI, 2020, available at https://hai.stanford.edu/news/geographic-bias-medical-ai-tools, which summarizes research showing that *"most [ML] algorithms [for clinical diagnosis tasks] are trained on datasets from patients in only three geographic areas, and that the majority of states have no represented patients whatsoever."* This is partly driven by the difficulty and expense associated with producing good datasets for training; developers often gravitate to using the data that is most readily available.

20 See, e.g., D. Ensign, S. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, *"Runaway feedback loops in predictive policing,"* Conference on Fairness, Accountability and Transparency, 2018.

21 See, e.g., D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, *"Machine Learning: The high interest credit card of technical debt,"* Google Research, 2014; E. Breck, S. Cai, E. Nielsen, M. Salib, D. Sculley, *"The ML test score: A rubric for ML production readiness and technical debt reduction,"* IEEE International Conference on Big Data, 2017; M. Zinkevich, *"Rules of Machine Learning: Best Practices for ML Engineering,"* Google Research, available at http://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf.

Ethics, governance, and policy

This section discusses a range of additional concerns about AI and ML, with a particular focus on ethics. It highlights several aspects of concern, including fairness and non-discrimination, and emphasizes public engagement as a key tool.

Ethics has become an increasingly prominent topic in AI in recent years. Work on this topic is often referred to as "AI Ethics," "responsible AI," or other similar terms. Not all these issues apply to every kind of AI application, and the discussion below is intended to give a feel for the topic, not to be fully comprehensive. It is worth emphasizing that all of the topics discussed here are the subject of active and recent research, so there is much that is not yet fully understood or settled.

It is above all important to make sure that the system in question actually works and accomplishes its goals. This is not as simple or as much of a given as it may sound.

Too often, systems are built around a flawed premise or simply do not work for their intended purpose; just because one can collect some training data and mechanically go through the motions of training a model does not mean it always makes sense to do so, or that the result should be taken seriously. Like many of the other potential failure modes described both above and below, this problem can manifest as anything from mere inconvenience or suboptimal performance to severe unethical behavior and impacts on real people and communities, sometimes including racism, sexism, or even matters of life and death.

To give a real example, a major electronic health record company sells an AI system for predicting whether patients will develop sepsis (a life-threatening condition that can arise in response to infection). This system is used by hundreds of hospitals around the country. In a recent study, researchers found that the model both performs substantially worse than the vendor reported, and poorly for clinical use in general: The tool misses two thirds of sepsis cases

"They took each other's advice, opened one book, went over to another, then did not know what to decide when opinions diverged so widely."

Gustave Flaubert
Bouvard et Pécuchet

(high false negative rate) while also overwhelming doctors with false alerts (high false positive rate).[22] This is also not one of the instances of discrimination that are discussed further below; the system simply does not work as it should across the board. Such examples underscore the need to rigorously evaluate system design and performance from a range of different perspectives.

While it is helpful to have an underlying ethical framework to underpin one's approach, it is still typically unclear how to actually operationalize such principles in practice.

When discussing ethics, it is important to think about what high-level ethical principles or framework are being used. In recent years, at least 100 institutions, from corporations to academics to nonprofits to national governments, have published various sets of "AI principles." In a comparative study, Harvard researchers found that there is broad overlap in these, which include privacy, accountability, fairness and non-discrimination, human control of technology, and others.[23] Though there is significant consensus around such principles at a high level, society and the field are still at the early stages of determining how to operationalize them. In the context of local governments, such principles have been referred to as "digital rights," by analogy with human rights, and these rights and principles are discussed in more detail below. For this reason, this document focuses more on practical challenges than on motivating or explaining the principles themselves.

AI forces developers, and society at large, to make societal and policy values and goals explicit.

To build an AI system, as described in the preceding sections, developers must specify things like what the system should be optimizing for and how different types of errors should be weighed. When building a system that impacts people, this often involves making ethical and other policy values quantitative and explicit. These values are not specific to AI: they are implied in any policy or decision-making process (such as hiring decisions, college admissions, or patient treatment) previously conducted solely by humans and governed by organizational policies. When used for

22 A. Wong, E. Otles, J. Donnelley, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, M. Ghous, and K. Singh, *"External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients,"* Journal of the American Medical Association–Internal Medicine, 2021.

23 J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, *"Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,"* Berkman Klein Center for Internet & Society, Harvard University, 2020, available at https://dash.harvard.edu/handle/1/42160420.

human decision-making, these typically were and continue to be implicit rather than explicit, and that implicitness in part can allow human decisions to remain unfair and inequitable.

As above, a mortgage loan AI system would need to specify the relative costs of wrongly denying a loan versus wrongly approving one, and possibly even explicitly break these down for different groups (by gender, race, or other attributes). This explicitness can cause discomfort, but it must be understood that only the explicitness, rather than the trade-offs themselves, are new, and making them explicit provides a significant collective opportunity to revisit and redesign how policies have been designed and implemented more broadly. Indeed, one of the potential uses of AI is to help inspect and evaluate how human decisions have been made.[24]

Accountability

Broadly, accountability in the context of AI refers to being responsible or answerable for the outputs, decisions, or impacts resulting from the use of an AI system or model.

This can take several different forms; the coverage here is not comprehensive but aims to give a feel for different ways in which this can be approached.[25]

One of the simplest forms of accountability is being transparent about the fact that an AI system is in fact being used to perform important functions or make impactful decisions. Although this may seem straightforward, it has been controversial in some contexts, such as the management of patient health.[26] Beyond this, one can consider providing transparency into specific aspects of the system, such as the data or models used.[27]

Another potential goal is to allow for some human intuition or understanding of what the model is doing, as opposed to just knowing how it performs on some validation data. There are different approaches to this, including just using much simpler models that are inherently easier to interpret, using additional technical methods

24 S. Mullainathan and Z. Obermeyer, *"Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care,"* National Bureau of Economic Research (NBER) Working Paper No. 26168, 2021; S. Mullainathan, *"Biased Algorithms are Easier to Fix than Biased People,"* The New York Times–Opinion, 2019, available at https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html.

25 In particular, see a recent GAO report on the federal government's approach to accountability, which *"identifies key accountability practices — centered around the principles of governance, data, performance, and monitoring"*: US Government Accountability Office, *"Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities,"* 2021, available at https://www.gao.gov/products/gao-21-519sp.

26 R. Robbins and E. Brodwin, *"An invisible hand: Patients aren't being told about the AI systems advising their care,"* STAT News, 2020, available at https://www.statnews.com/2020/07/15/artificial-intelligence-patient-consent-hospitals/.

27 Some of these ideas have informally been referred to as "nutrition labels" for ML; see, e.g., T. Gebru, J. Morgenstern, B. Vecchione, J. Vaughn, H. Wallach, H. Daumé, and K. Crawford, *"Datasheets for datasets."* arXiv preprint arXiv:1803.09010, 2018; M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, D. Raji, and T. Gebru, *"Model cards for model reporting,"* Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019.

to "inspect" the inner workings of more complex models, or designing the system to allow users to see how the model's outputs vary as certain input attributes are changed.[28] These sorts of decisions may allow one to see that, for example, borrower age is or is not very relevant in predicting loan approval, or indicating to a borrower or loan officer that a loan would have been approved if the applicant's income were over some number. Which approach makes sense, and is amenable to different stakeholders, will depend on context; the expectations and resulting approach will be different across consumer finance, medical diagnosis, and criminal justice, and are still subject to ongoing research and debate.[29]

When humans are involved, they need to be considered part of the system itself.

A different way to approach accountability is to integrate human oversight via maintaining a role for humans in the ultimate decisions. For example, the system may only make suggestions or help focus the human's attention on the more ambiguous or difficult cases. These are sometimes referred to as "partially automated" or "human-in-the-loop" systems. In these cases, the way in which the human operators are trained to use, override, or ignore the system, as well as how the interfaces of the system are designed, play a critical role in overall system behavior. In addition, the way the system's suggestions are framed, described, or presented can have an outsized impact on how the human in question reacts to them. For example, a user may interpret a "green light/red light" display very differently than a risk assessment score displayed with five decimal points, and such seemingly minor details can in turn influence, sometimes dramatically, the ultimate behavior of the "whole system." For this reason, it is critical to engage experienced designers when building systems that include human interfaces.[30]

For example, if human operators are allowed to override or deviate from the recommendations or decisions of a mortgage loan model, what matters and must be evaluated is whether the overall mix of computer and human decisions satisfy the desired goals, rather than the model by itself in a vacuum.[31]

28 Some of these different topics are referred to as interpretability, explainability, or transparency.

29 See, e.g., C. Rudin, *"Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,"* Nature Machine Intelligence, 2019; K. Miller, *"Should AI Models Be Explainable? That depends,"* Stanford Center for Human-Centered AI, 2021; B. Haibe-Kains et al, *"Transparency and reproducibility in artificial intelligence,"* Nature, 2020; H. Stower, *"Transparency in medical AI,"* Nature Medicine, 2020; J. Vaughn, *"Transparency and Intelligibility Throughout the Machine Learning Life Cycle,"* available at https://www.youtube.com/watch?v=I-TSjiXGfSI.

30 See, e.g., M. K. Lee, D. Kusbit, E. Metsky, and L. Dabbish, *"Working with Machines: The Impact of Algorithmic, Data-Driven Management on Human Workers,"* ACM/SIGCHI Conference on Human Factors in Computing Systems, 2015; E. Tufte, *Visual Explanations,* Graphics Press, 1997; D. Huff, *How to Lie with Statistics*, Norton, 1954.

31 This is sometimes referred to as a "sociotechnical systems" approach.

Fairness

The term "fairness" in AI refers to the notion that systems should not discriminate with respect to certain personal attributes or protected characteristics, such as race, gender, age, or disability.

This can take different forms depending on the situation. For example, it may be that loan applications that are otherwise similar are declined at much higher rates for women than men,[32] or that a system is much less accurate for certain groups of people in a way that results in some kind of disparate impact or harm.[33] This topic is an area of significant concern and has received a great deal of attention in recent years.[34]

Though this behavior is sometimes referred to as "bias" or "algorithmic bias," this document uses "fairness" both to avoid confusion with other unrelated technical definitions of "bias" in AI, and to emphasize that the concern is ultimately about impacts on people rather than a narrower consideration of model behavior alone.[35]

For example, a model that by itself is "unbiased" in some technical sense can turn out to have unfair outcomes for people when actually deployed (sometimes because of decisions humans-in-the-loop make outside the model itself). On the flip side, it may be possible to use a model that is technically "biased" to promote equity and advance other goals.[36] For example, in some preliminary work, researchers partnering with the Los Angeles City Attorney's office found that they could have a possibly biased system result in equitable criminal justice outcomes across racial groups by, among other things, coupling technical considerations with a tailored social service intervention strategy. [37] Alternatively, in the mortgage context, a model may be designed to support lending decisions in a way that actively corrects historical racial disparities in homeownership rates.

In sum, narrowly focusing on developing AI models and algorithms that better account for fairness will generally not be sufficient to actually achieve more equitable decisions or outcomes, which is the real goal. Instead, efforts should work towards making

32 See, for instance, L. Goodman, J. Zhu, and B. Bai, *"Women Are Better than Men at Paying Their Mortgages,"* Urban Institute –Housing Finance Policy Center, 2016.

33 S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities,* 2021, accessible at http://www.fairmlbook.org.

34 For overviews, see, e.g., J. Vaughn and H. Wallach, *"Machine Learning and Fairness,"* 2020, available at https://www.youtube.com/watch?v=7CH0xLWQLRw; H. Wallach and M. Dudik, *"Fairness-related harms in AI systems: Examples, assessment, and mitigation,"* 2021, available at https://www.youtube.com/watch?v=1RptHwfkx_k; K. Rodolfa, P. Saleiro, and R. Ghani, *"Bias and fairness,"* chapter of *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*, available at https://textbook.coleridgeinitiative.org/chap-bias.html.

35 See, e.g., J. Buolamwini and T. Gebru, *"Gender shades: Intersectional accuracy disparities in commercial gender classification,"* Conference on Fairness, Accountability and Transparency, 2018.

36 R. Ghani, *"Equitable Algorithms: Examining Ways to Reduce AI Bias in Financial Services,"* Testimony to Artificial Intelligence Task Force, Committee on Financial Services, U.S. House of Representatives, 2020.

37 K. T. Rodolfa, E. Salomon, L. Haynes, I. Mendieta, J. Larson, and R. Ghani, *"Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions,"* ACM FAccT, 2020.

entire systems — including the humans and organizations involved — and their ultimate outcomes and effects fair.[38]

There is no single definition of fairness and the notion and goal appropriate for the situation at hand must be determined through the development process; however, it is often not appropriate for the developers to make these decisions unilaterally, so broader stakeholder engagement is often necessary.

In some instances, there may be laws requiring that certain types of decisions are made fairly, including definitions of what is meant by "fair" in that domain; examples include the Fair Housing Act, the Equal Credit Opportunity Act, or the Uniform Guidelines on Employment Selection Procedures. In other cases, it may be necessary for the developer to choose what "fair" should mean, and there are a large set of formal criteria with different implications.[39]

In the context of lending,[40] one possible criterion is "race and gender blindness"; in other words, a requirement that the system not include race or gender as input features. This is straightforward to implement, but is not likely to actually avoid disparate outcomes, partly because other input features can serve as proxies for race or gender.[41] In particular, because of the historical effects of redlining, a seemingly simple piece of information like ZIP code will often serve as a proxy for race and lead to the system being racially unfair. In addition, not letting the model see this data may prevent the developer from using certain types of technical corrections to ensure fairness across those attributes.

A different definition could be "parity," meaning that the exact same number (or percentage) of loans should be approved or denied in each demographic group. On the one hand, this may ensure that protected groups would get loans; on the other, it may mean they are more often getting loans they cannot repay. Yet another definition might be that the decisions serve to reduce disparities in homeownership rates across racial groups. In short, there are dozens of ways one might define fairness, and these conflict with each other in that a system will be fair under one definition but unfair by another. Ultimately, deciding on an appropriate standard

38 For a discussion of practical implementation issues in fairness, see, e.g., C. Bakalar, R. Barreto, S. Bergman, M. Bogen, B. Chern, S. Corbett-Davies, M. Hall, I. Kloumann, M. Lam, J. Candela, and M. Raghavan, *"Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems,"* arXiv preprint arXiv:2103.06172, 2021.

39 A. Narayanan, *"21 fairness definitions and their politics,"* Conference on Fairness, Accountability, and Transparency, 2018; S. Verma and J. Rubin, *"Fairness definitions explained,"* IEEE/ACM International Workshop on Software Fairness, 2018.

40 For a summary of several different fairness criteria for mortgage lending, see, for example, https://www.aspentechpolicyhub.org/wp-content/uploads/2020/07/FAHL-Cheatsheet.pdf.

41 See, e.g., Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, *"Dissecting racial bias in an algorithm used to manage the health of populations,"* Science, 2019.

is a context-specific policy decision that cannot be made on purely technical grounds, and needs to incorporate interdisciplinary expertise and values.

The root causes of model malfunctions that lead to unfair outcomes (even impacts that are disparate by race) often have nothing directly to do with race or other demographics.

Importantly, the types of effects above, though undesirable, are in no way specific to protected characteristics like race. For example, an AI system for medical diagnosis (say, determining whether an MRI scan contains a tumor or not) may be more accurate using images from Manufacturer A's hardware than it is using images from that of Manufacturer B. In itself, this has nothing to do with demographics. However, if the hospitals using Manufacturer B's hardware were resource-constrained and served a poorer patient population, this difference could serve as a proxy for class or income, and then the overall system may end up producing disparate impacts when actually deployed - and, for example, over- or under-diagnosing tumors at much higher rates for certain groups. In addition, lower-income populations tend to disproportionately include people of color, women, and other protected groups, so this differing performance by manufacturer is likely to produce other disparate impacts as well, even though this is far from the root "problem" in the system. But even if none of this were the case, this kind of system would still pose a serious concern, as all patients, including those not in protected classes, could be harmed by a model that is not performing well on their local equipment.

For this reason, the overall potential impacts of a system must be considered in evaluating whether a system is functioning appropriately or not.

The effects described above can arise at any stage of the AI lifecycle, including problem formulation, data collection and processing, and modeling. In some cases, the system's malfunction may be for straightforward reasons; for example, the medical diagnosis system may be inaccurate for images from Manufacturer B's hardware simply because there were not enough such images included in the

training set. In this case, the solution may be as simple as getting more such images and retraining the model.[42] In other cases, the problem may be more subtle and not simply related to data.

Ultimately, this is a complex and evolving topic with no simple answers, much like most policymaking. Having said this, developers should become aware of some of the sources of unfairness and proactively consider them. For example, race, gender, class, and other factors are inextricably tied to data in many social and economic domains, including geographic and housing data, medical treatment, consumer and business finance, employment, and criminal justice. One would need to be much more vigilant in these areas and with such data than, for example, ML models used to help manage battery usage in a phone.

One example of a concrete way to identify potential issues is to carry out so-called "disaggregated evaluation," or an evaluation of model performance broken out by different subgroups in the data, demographic or otherwise.

This can reveal, for instance, if a model is performing well for the population overall but very poorly for some group that is a small percentage of the data. Although there are valid legal privacy concerns about the collection of sensitive data like race, gender, and disability, it is important that corporations and governments are able to carry out these sorts of analyses in order to ensure fairness or equity.[43]

Privacy and security

Privacy and cybersecurity are two of the most important digital rights topics, and they also apply to AI and ML. In some cases, these apply to the collection and use of data in general; for example, some data is considered more sensitive, either because it is "identifying" or because it concerns a sensitive topic, such as medical information about an individual.[44] In these cases, there may be tighter restrictions around the use of this data and more rigorous expectations of how the data should be protected. These are

42 See, e.g., A. Kaushal, R. Altman, and C. Langlotz, *"Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms,"* Journal of the American Medical Association, 2020.

43 See, e.g., S. Barocas, A. Guo, E. Kamar, J. Krones, M. Morris, J. Vaughan, D. Wadsworth, and H. Wallach, *"Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs,"* arXiv preprint arXiv:2103.06076, 2021.

44 A detailed discussion of terms like "identifying" is out of scope here, but a core concept in privacy law is "Personally Identifiable Information" or PII. There is no standard definition of PII, and particular policy frameworks and jurisdictions have their own definitions. For some general discussion, see P. Schwartz and D. Solove, *"The PII problem: Privacy and a new concept of personally identifiable information,"* NYU Law Review, 2011.

illustrated in traditional privacy frameworks like the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which includes a Privacy Rule (which defines notions such as Protected Health Information) and a Security Rule (which defines security expectations around electronic records). These topics are not discussed further here, as they are not specific to AI.[45] This section instead briefly highlights some other aspects of these issues.

Although more general issues about data privacy and cybersecurity also apply in the context of AI, there are also novel privacy and cybersecurity concerns that arise.

Some of these topics are technical and not discussed in detail here, but briefly, there are specialized privacy attacks that can apply to ML models. Two examples related to privacy are "membership inference" and "model inversion," which broadly involve learning things about the underlying data that was used to train a model given access only to the trained model itself.[46] When a model is used in a sensitive domain, this may be a concern. Similarly, "proxies" (discussed above) can implicitly introduce a form of privacy loss. In security, there are concerns about "software supply chains": because ML relies heavily on a shared set of resources in the form of datasets, models, and software libraries (which themselves depend on other, lower-level libraries), many of which are open source, there are questions about their vulnerability to digital supply chain attacks.[47] These examples are merely illustrative.

Privacy, security, fairness, accuracy, and other desirable goals or characteristics of systems are often in tension with each other. The trade-offs between these principles should be explicitly acknowledged, and developers must proactively and explicitly determine how best to navigate these trade-offs in any specific situation. This may require input from a range of stakeholders.

The issue of trade-offs between different aspects of AI systems, especially various digital rights, is one of the most fraught in the field.[48] For example, there can be trade-offs between privacy and accuracy (or other measures of model performance). Usually, the more data is used, the better the model will perform. This can

45 See P. Ohm, *"Broken promises of privacy: Responding to the surprising failure of anonymization,"* UCLA Law Review, 2009, for an overview of some recent issues in modern information privacy.

46 R. Binns, *"Privacy attacks on AI models,"* UK Information Commissioner's Office, 2019, available at https://ico.org.uk/about-the-ico/news-and-events/ai-blog-privacy-attacks-on-ai-models/.

47 A. Lohn, *"Poison in the Well: Securing the Shared Resources of Machine Learning,"* Georgetown Center for Security and Emerging Technology, 2021, available at https://cset.georgetown.edu/publication/poison-in-the-well/.

48 R. Binns, *"AI Auditing Framework: Trade-offs",* UK Information Commissioner's Office, 2019, available at https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/.

include collecting or using data on a larger number of people or augmenting data about each person with additional demographic or other data, both of which could be determined to reduce how "privacy-respecting" the system is. On the other hand, avoiding collecting demographic data, including data that is sensitive, may degrade accuracy, and could result in faulty output or even harmful consequences (in, say, an inaccurate medical diagnosis).

Similarly, privacy and fairness can conflict in different ways. In one case, if developers find that its system is unfair due to insufficient training data on a particular demographic population, they may want to collect more data from such groups to increase model accuracy. Separately from this, in order to test whether an AI system is unfair or discriminatory in the first place, it is generally necessary to collect data on populations with protected characteristics (e.g., to carry out disaggregated evaluation, as described above). The developers would then face a trade-off between privacy (not collecting the data on characteristics) and fairness (collecting and using the data to test the system and make it fairer). This trade-off is not theoretical; indeed, lacking access to this data is cited by practitioners as one of the chief impediments to building fairer AI systems.[49]

This is just one example; there are a range of other ways to balance privacy protections while enabling productive use or sharing of data, such as through synthetic data, de-identification, privacy-preserving data analysis algorithms (e.g., using secure multiparty computation), or governance structures like data sharing or confidentiality agreements where law permits the data to be shared.

There can also be trade-offs between accuracy and fairness, privacy and data security, explainability and accuracy, and so on. All in all, none of these rights can be taken to be a universal good.

Community engagement and participation

Public engagement is the work done by officials to meet and invite constituents into the processes of governance. This work takes many forms, from town halls and community boards to new

49 K. Holstein, J. Vaughn, H. Daumé, M. Dudik, and H. Wallach, *"Improving fairness in machine learning systems: What do industry practitioners need?,"* CHI, 2019.

technology-enabled modes of engagement such as crowd-sourcing apps, participatory democracy platforms, or social media. No matter the form, public engagement is guided by the democratic principle that decisions should be made with the public, not just for the public. This is especially important in contentious areas of public life with high stakes, such as public safety, public health, education, or child welfare.

Such engagement is also important in AI and can be an essential aspect of the responsible development, use, and governance of certain systems. In the public sector, it is particularly important for the public to be engaged because the government is responsible for ensuring that technology reflects the concerns, needs, and values of constituents, accurately accounts for impacts, and is deployed in an accountable manner, ideally in a way that supports a sense of trust, respect, and empowerment among constituents.

Especially because this is an emerging topic in the context of AI, determining when and how to do engagement, and what form it should take in each given situation, can be complex and challenging; best practices and standards for robust public engagement in AI are not yet agreed upon and are themselves the subject of active current research, and the complexity and novelty of these systems will likely require new, innovative methods to enable robust and meaningful participation.

Engagement plays an essential role for several practical reasons in addition to the high-level principles above. For example, community concerns that arise in engagement efforts can sometimes overshadow any benefits the system may have for the community involved, and when AI systems are deployed that do not reflect community needs, either in actuality or in perception, they may receive pushback from the community and ultimately not be adopted regardless of any other merits of the project or how careful the developers may have been with considering other ethical issues. There are many examples of this around the world, from residents organizing against the installation of biometric locks in their housing complexes to public protests over the way models

were used to guess at what scores students were expected to get on admission exams.

There are also many applications that do not involve direct or meaningful human impacts — such as the algorithms used to optimize battery usage in a smartphone or many internal administrative applications in organizations — and in those cases, engagement may be unnecessary or can even be counterproductive.[50]

50 For administrative applications, attention to topics like impact on work, organizational structures and processes, and job security can also be important.

Public engagement should be considered with any system or process that uses computation to aid or replace decisions or policies that impact opportunities, access, liberties, rights, or safety.

Engagement can be used for different purposes and at different points in the AI lifecycle, depending on the context of the project. In the following example, the engagement was done before system deployment and to help design the system itself; in other cases, engagement has been done post-deployment.[51]

51 See, e.g., A. Brown, A. Chouldechova, E. Putnam-Hornstein, A. Tobin, and R. Vaithianathan, *"Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services,"* Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019.

412 Food Rescue and participatory system design

412 Food Rescue is a non-profit located in Pittsburgh that matches donor organizations with expiring food to non-profit recipient organizations, and the organization decided to build an AI system to allocate donations because its existing manual approach was both time-consuming and inequitable. However, a difficult trade-off quickly arose: because the donors tended to live in different and wealthier areas than recipients, increasing equity (allocating to recipients with greatest need) meant decreasing efficiency (longer distances to travel for volunteers).

412 partnered with researchers at Carnegie Mellon University to develop the system in a participatory way.[52] Researchers had stakeholders participate in each stage of the AI development lifecycle, including determining which input features they felt were important (such as travel time, income level, or food access), voting on which model predictions best reflected what they would consider an appropriate equity-efficiency trade-off, and more. Importantly, the process of "engagement" or "participation" is not simply a matter of hearing people's opinions; here, it involved implementing a

52 M. K. Lee, D. Kusbit, A. Kahng, J. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, and A. Procaccia, *"WeBuildAI: Participatory framework for algorithmic governance,"* ACM: Human-Computer Interaction, 2019.

structured way to elicit data on people's preferences as well as a rigorous method (in this case, based on the theory of social choice from economics) to aggregate individual opinions into an overall policy. When stakeholders were interviewed after system implementation, researchers found that they felt the system was fair, in part due to the participatory approach used and in part because of the actual outcomes achieved. In this case, stakeholders needed to be engaged before the system was deployed; although many social domains exhibit complex trade-offs of this type in which there is no obvious "right" answer, it can be possible to directly incorporate varied views in the design to both achieve better outcomes and earn the trust of the people involved.

Conclusion

To build a healthy AI ecosystem for New York City, local decision-makers must work with a clear understanding of the technology and key practical and ethical considerations around its design, use, and governance. This AI Primer aims to equip decision-makers with a helpful foundation as they begin to engage with AI in an increasingly broad range of ways — to build teams, evaluate products, outline governance measures, formulate policy, or simply work to further develop their own knowledge and capacity.

Using and assessing AI can be a highly complex endeavor, and each case requires a detailed evaluation of a range of goals and factors. These can both be in tension with each other and be highly contingent on the context of each case. In that sense, this Primer can serve as an aid, but there can be no explicit prescription for how to make sound decisions, just as there can be no universal formula that guides policymakers when navigating difficult trade-offs. Indeed, as we have emphasized, many key issues that arise will often not be about AI itself or technical details at all. Moreover, this field, and many of the particular aspects described here, are very rapidly evolving, and it is not uncommon to see research or reporting that upends the status quo in a given area. The breadth and range of teams and use cases is further fueled by the fact that access to this technology, even at relatively large scale, is increasingly widely available and not limited solely to large corporations or governments, or even to individuals with significant technical expertise or formal training.

In AI, more so than many other areas, there are still far more questions than answers. For all these reasons and others, it will be critical to continue to learn and to evolve this framework as these broader efforts progress.

Further references

For some common texts on machine learning, see C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006; T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning,* second edition, Springer, 2016; I. Goodfellow, A. Courville, and Y. Bengio, *Deep Learning*, MIT Press, 2016; M. Hardt and B. Recht, *Patterns, Predictions, and Actions: A story about machine learning*, 2021, available at https://mlstory.org.

For a more general textbook on AI, see S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, fourth edition, Pearson, 2020.

There are also books focused on implementation, such as F. Chollet, *Deep Learning with Python*, O'Reilly, 2021; A. Géron, *Hands-On Machine Learning with scikit-learn, Keras, and TensorFlow*, O'Reilly, 2019; J. Howard and S. Gugger, *Deep Learning for Coders with Fastai and PyTorch: AI Applications Without a PhD*, O'Reilly, 2020.

For a general perspective on these areas, see M. Jordan, *"Artificial intelligence—the revolution hasn't happened yet,"* Harvard Data Science Review, 2019, and M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*, Farrar, Straus and Giroux, 2019.

fin.