USING ARTIFICIAL INTELLIGENCE AND BIG DATA TO DEVELOP TOOLS USED IN COURTS

By Alan Carlson*

version 5, November 11, 2019

Table of Contents

* Engineer and lawyer; retired as Court Executive Officer and Clerk of Court, Orange County (CA) Superior Court.

INTRODUCTION

(1) Serious attention is now being focused on using Ai to develop tools that assist people in navigating legal processes, making legal decisions, or resolving legal disputes. Using artificial intelligence (AI) to analyze available datasets could greatly increase the availability, effectiveness, and usability of such tools. At the same time, there are risks of harm if these tools are not developed and used responsibly. The potential of harm is particularly serious in the legal setting where a tool's predictions can directly affect

critical aspects of life, for example, incarceration or loss of property or other rights, as opposed to shopping recommendations. In light of the potential harms, the objective of this article is to identify risks, discuss their sources, and propose ways to mitigate the risks. Identifying and addressing risks will allow the development of more effective and responsible AI developed tools.

(2) The intended audience of this paper are people looking into the use of AI developed tools in courts. The discussion is intended for policy makers (judges, legislators, and court and public agency managers), as well as tool developers because it is policy makers, not tool developers, that should, at the outset of development, define goals and establish accountability for any AI developed tools.

(3) The discussion is not a "how to" manual for developing AI tools. Nor are the differences and nuances of various AI algorithms that could be used in a court setting explored. Rather, the discussion provides a general, non-technical overview of benefits and risks of using AI developed tools. Although the discussion identifies risk mitigation approaches, it does not provide specific technical solutions. The field of AI is developing rapidly, generating new approaches and solutions beyond those noted here. But, to start a journey, one must begin from where they are.

TERMINOLOGY

(4) The fields of artificial intelligence and machine learning use a lot of technical terms, many carried over from mathematics and statistics. Rather than try to introduce and use a lot of technical terms, simpler, more generic, terms are used. The point is to introduce the subject to non-technical readers. There are a few key terms used that are defined as follows for purposes of this paper:

(5) The term **AI** is used here to refer to any analytic, mathematical, or statistical method or algorithm used to develop a tool that assists people in navigating the judicial process, making legal decisions, or resolving legal disputes. AI, as used here includes machine learning, deep learning, neural networks, reinforcement learning, supervised and unsupervised learning, expert-system, decision theory, data mining, data analytics, etc. Tools based on AI have been developed for a variety of applications, including visual perception, image recognition, speech recognition, translation, interpretation, prediction, classification, and decision making. This paper focuses on decision-making applications in a legal setting.

(6) **Tool** refers to a model, agent, system, algorithm, product, service, app, etc., developed using AI that is used by someone in a legal setting to navigate the judicial process, to help make legal decisions, or to resolve a legal dispute. The tools contemplated here involve decisions about human beings, not tools controlling mechanical or computer based systems.

(7) **Operator** refers to the person or entity using an AI developed tool to make or support a legal decision. An operator applying a tool can be a lawyer or a staff person in a law office (private, public, or non-profit). It could also include a judge, or a staff person working for the court or for the clerk of court, or for a government agency supporting the legal system, for example, providing pretrial release, probation, or detention services.

(8) **Subject** refers to the person or entity to whom the tool will be applied. The subject can be a litigant or a party to a case, a potential litigant, or a member of the public. A person applying a tool to themselves is both an operator and a subject.

(9) A **Training Dataset** is the set of data elements about examples or scenarios that are used to 'train' a tool in making predictions, classifications, or decisions. Each example or scenario typically consists of several data elements (sometimes referred to as inputs, values, variables, attributes, features, etc.). For some AI algorithms, each example in the training dataset must also include the outcome, or label, the tool should predict for that example. For other approaches this is not necessary.

AI TOOL USE CASES IN THE JUDICIAL SYSTEM

(10) To provide a context for the discussion of benefits and risks associated with AI developed tools, it is useful to have examples of how AI developed tools are, and could, be used in the justice system. The following are examples of existing or potential AI tool use cases, including some law office applications that may also be used in the justice system. Most are applications already in production, a few are applications being tested or deployed in a few jurisdictions, and some are just concepts for applications.

(11) *On-line legal research*. Natural language search methods are used to identify what law (cases, statutes, regulations, scholarly analysis, practice guides, etc.) might be applicable to a specific legal issue. Text search tools have been 'fine-tuned' to use phrases and citation information as well as basic word search terms.

(12) *Automation of routine tasks*. Extraction of data from court documents to: a) provide data entry to an electronic filing system or a case management system, b) fill out forms, c) take action based on the value of data extracted (for example, identifying documents or data as confidential or to be sealed), or d) assign cases to case management tracks based on the nature and complexity of the case. This approach is called Robotic Process Automation (RPA).

(13) *Document assembly*. Generation of legal documents from templates using data in a case management system or using responses to a questionnaire-style interview. This approach is now commonly used to assist self-represented litigants.

(14) *Risk Assessment.* Pretrial release tools that assist judges in deciding whether to release a defendant who has been arrested pending trial. A few jurisdictions are also using risk assessment tools for sentencing decisions.

(15) *On-line dispute resolution*. Several tools are available that assist people in resolving disputes on-line, without 'coming to court'. Some are very subject specific, for example, for traffic citations, others are more generally applicable. Many are focused on smaller, but high-volume court cases, such as traffic citations, landlord-tenant/eviction, or small claims cases.

(16) *Document review*. Another use case is the review of legal documents. One example is review of a proposed order in a family law case. The review could assess: a) the compliance of a proposed order with applicable law, b) completeness, and c) consistency with generally accepted practices, for example, child custody arrangements. It could also flag potentially unjust orders, or an imbalance of provisions based on an asymmetry of power between the litigants.

(17) *Producing the verbatim record* directly from an oral hearing or recording of an oral hearing. A tool could use speech to text conversion to generate a written transcript of the oral testimony given at a hearing or the judges verbal order after hearing.

BENEFITS OF USING AI TOOLS

(18) There are many potential benefits and advantages in using AI developed tools to guide people through the resolution of legal disputes and assist them in making legal decisions. The following list highlights ways in which AI developed tools could be of benefit.[1] Not all of these would require tools using the more sophisticated AI algorithms and some may be possible using simpler analytic tools. The point is to summarize the ways in which tools can both benefit operators and subjects and improve the efficiency and effectiveness of the legal process.

(19) IN <u>BUILDING</u> A TOOL OR WEB PORTAL, AI developed tools can:

a) Assist in the initial design of a web portal by providing information about:
   (i) What people's legal problems are;
   (ii) What alternatives and outcomes are possible or appropriate for these legal problems, possibly including some without 'going to court'; and
   (iii) What the alternative pathways in a web portal guiding people through the legal process should be.
b) Assist in maintaining and upgrading a web portal by analyzing information about people's experiences using the web portal or analyzing new datasets thereby suggesting what the next version of the web portal might include. It can also assist development of web portals for other legal issues through a transfer of knowledge about how to build a web portal.

(20) IN USING A TOOL OR WEB PORTAL, AI developed tools can:

a) Improve the subject's experience through greater transparency regarding the legal process and the predictions offered that will allow subjects to make more informed decisions;
b) Identify or flag issues relevant to the resolution of a subject's legal issues or dispute;
c) Assist the subject in navigating through the legal process and web portal pathways;
d) Provide information to the subject about the range of reasonable outcomes in a given context;
e) Offer available options or solutions to the subject, including alternatives to traditional adjudication;
f) Suggest most likely outcomes and help a subject choose among the options available;
g) Provide referrals to context-appropriate service providers, such as mediators, on-line dispute resolution providers, and social or mental health services or providers;
h) Increase the use of web portals because of the assistance and opportunities they offer to address people's legal problems; and
i) Help restore public trust and confidence in the judicial system's responsiveness and ability to help people resolve legal disputes.

(21) IN SUPPORTING LEGAL DECISION-MAKING, AI developed tools could:

a) Provide information to the subject about all of the outcome patterns in the dataset used to develop the tool;
b) Provide information to the subject about the most likely outcome in a given situation based on past practices and experiences;
c) Significantly shorten the time needed to gather the most directly relevant information and to make an informed decision;
d) Help the subject avoid unconscious or implicit bias or confirmation bias often present in human decision making; and
e) Detect and incorporate relevant factors which human decision makers have overlooked, were unaware of, or for which they underestimated the relevance or significance.

(22) OVERALL, AI developed tools should:

a) Improve the efficiency of the legal system for operators and subjects;
b) Improve the overall efficiency of the legal system;
c) Improve the transparency of the legal system processes;
d) Increase public trust and confidence in the legal system; and

e) Assist in the redesign of legal processes.

BUILDING AN AI TOOL

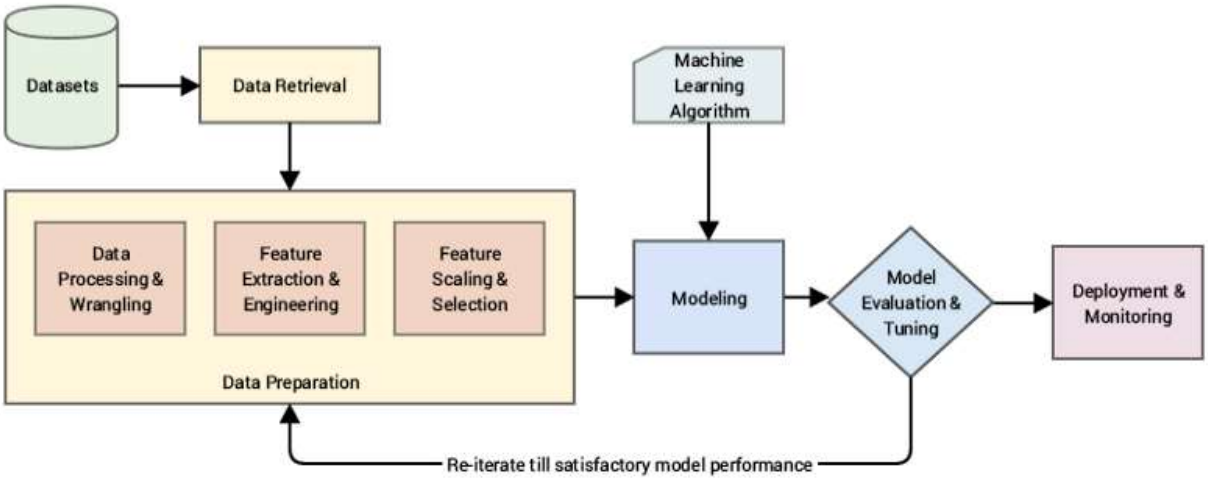
(23) In order to better understand potential risks from using AI developed tools and what can be done to mitigate them, it is helpful to understand both how tools are typically developed and how they 'work'.

Process of Building a Tool

(24) The critical first step when contemplating the development of a tool is to clearly and explicitly articulate the goals and business objectives the tool is intended to support.[2] At this stage the project should articulate a clear problem statement – what problem the tool will solve – and what outcomes or predictions are sought and under what conditions. This both guides the tool development process and provides a realistic check as to whether the tool is a success.

(25) The next set of decisions is about performance. How accurate do you want your tool to be? What is an acceptable level of precision, given that the tool's predictions are statistical, not deterministic? This involves both the choice of metrics, and the values of the chosen metrics that the tool should achieve in order to be considered successful.[3] There are often tradeoffs between precision and accuracy. In some situations the tool may need to be more precise – minimize the proportion of wrong predictions. In other situations, the tool may need to maximize the proportion of correct predictions. Having the right performance metrics in light of the tool's goals is critical in evaluating a tool's success.

(26) Once a decision is made to develop a tool, and the goals and performance metrics have been articulated, the typical steps in building a tool are shown in the following graphic.[4]



Source: Practical Machine Learning with Python, Apress/Springer

(27) Having articulated an understanding of what the tool is supposed to predict, data thought to be relevant to the desired prediction is collected for a set of examples or scenarios similar to those to which the tool will be applied. [5] Identifying relevant data points, gathering the data, cleaning it, and transforming it (referred to as feature engineering), often takes more time than developing the model itself. The bulk of the data collected becomes the training dataset, with a portion of the initial training dataset set aside to serve as a "test dataset" for evaluation of a model once it is built.

(28) After the training dataset is prepared, one or more AI algorithms are applied to the training dataset to build a model. Broadly speaking, there are three types of algorithms for building a model. One type, referred to as an expert-system, is to articulate all of the rules and decision logic the tool is to use in making a recommendation. Identifying and stating all the rules and decision logic can be a daunting task even for a modestly complicated system, let alone the legal system. Nevertheless, if the rules and decision logic are explicit, the inner workings of the tool are relatively transparent to anyone who seeks to 'examine the code' to know how the tool reached the recommendation it made.

(29) Another set of algorithms 'learns' by observation and generates a tool that produces a prediction based on the pattern of practices contained in the training dataset. Stated simply, this approach identifies patterns from examples and builds a model that is then used to make a prediction for a new example to which the tool is applied. Many of the risks and mitigation measures discussed below apply primarily to the tools developed using this approach. If the learning is based on examples in a training dataset where the predicted outcome for each of the examples in the dataset are provided, it is referred to as "supervised learning". If the learning does not depend on labelled outcomes, it is referred to as "unsupervised learning".

(30) A third set of algorithms 'learns' by repetition, either from continuous analysis of information 'fed' into the analysis or from generating numerous scenarios based on the applicable 'rules' and identifying successful strategies for making predictions. DeepMind's AlphaZero 'learned' to beat humans at the game of Go by first playing millions of games with itself.[6] It is unclear what the analogous rules would in a legal setting to build an 'AlphaGo' type tool for legal decision making.

(31) Once an algorithm is selected, there will generally be many iterations against the training dataset until the model achieves the chosen value for the performance metric. Once a model achieves acceptable performance, it is applied to the test dataset to further assess the accuracy of predictions. If the model does not exhibit acceptable performance on the test dataset, more iterations or other modifications of the model are needed. Note that there should be no expectation that the model developed will be 100% accurate in its predictions. Most AI algorithms are based on statistical correlations, so the predictions made are those that are statistically more likely. It is

possible that a tool that is sufficiently accurate in making predictions for a training dataset may not achieve the same level of accuracy when applied to the test dataset or the real world.

(32) Once a model provides satisfactory performance, it can be deployed and applied to new scenarios. With some AI techniques, the model is 'fixed' and does not change based on new scenarios to which it is applied. Other models can continuously learn as it is applied to new data (referred to as continuous integration).

Characteristics of Tools Developed Using Ai

(33) An AI developed tool does not know everything – it only knows what is in the training dataset used to build it. If information relevant to the best prediction is not present in the training dataset, then it cannot become part of an AI tool's predictions. Conversely, if data is included in the training dataset which is not related to, or is irrelevant to, the predictions sought, the tool may find patterns in this irrelevant data and use those patterns to make predictions which are non-sensical to the prediction sought.

(34) AI approaches involve the analysis of a training dataset containing information about past events and outcomes of the same or similar type as those to which the tool will be applied. For example, a tool to assist in making a pretrial release decision is built based on a training dataset of past pretrial release decisions. Since the training dataset contains only past decisions, the tool will necessarily generate predictions consistent with past decisions. Moreover, AI approaches produce tools that optimize predictions of the existing practices contained in the training dataset. While AI identifies existing patterns and practices, perhaps finding some that humans may not have consciously been aware of, AI cannot discover new ways of doing things, improvise, or offer other innovative practices or outcomes that are not present in existing practices.

(35) Accountable deployment and use of a tool requires that the scenarios to which the tool is to be applied to be essentially similar to the scenarios in the training dataset. AI approaches cannot anticipate, or change in response to, shifts in the scenarios to which it is applied, unless it is continually updated with data about the new scenarios. If a tool is applied to new scenarios not present in the training dataset, it should not be expected to make useful predictions.

(36) The mechanism used by many AI algorithms is to roam through a training dataset looking for patterns. Currently there is no capacity in AI's roaming to incorporate in its analysis the context of the data points or what the underlying principles are that generated the combination of data points the analysis is roaming. For an AI analysis the quantity of data points is more important than the connections or relationship between data points. Just because a model uses certain data points to make a prediction, it does not necessarily follow that the data and prediction are actually related – statistical

correlation does not imply causation. The patterns used for predicting may be merely coincidental, or may exist in the training dataset but not in the 'real world'.

(37) AI developed tools approach problems differently than humans. When addressing a problem and making a recommendation, human minds can apply many types of thinking: deduction, induction, symbolic reasoning, spatial logic, emotional intelligence, short-term memory, long-term memory, heuristics, etc. In addition, humans rely on the collective learning of other humans – a "society of minds".[7] Although there are many AI approaches to problems, the underlying processes are not as varied as human processes are. And AI cannot, yet, incorporate the learning of other AI developed tools the way humans can incorporate the learning of other humans.

(38) The current state of many AI algorithms is such that it is generally impossible to explain how or why the tool reached the prediction it offers. This is more true of some AI algorithms, for example, deep reinforcement learning. This is more difficult when an AI developed tool is built using a training dataset that contains a large number of data variables for each example. With a large number of inputs, it may not be readily apparent how many factors or which were the most important or relevant to the prediction made regarding a particular scenario.

RISKS IN USING AI DEVELOPED TOOLS

(39) There are several types of risks involved in using AI to build tools to navigate the legal process, to help make legal decisions, or to resolve legal disputes. Which risks are most problematic depends on what a tool is designed to do, what algorithms it uses, and what data was used to train it. Not every risk discussed here is inherent for every kind of algorithm.

(40) Twelve categories of risk are identified below with a brief exploration of the nature of each risk, the common sources of the risk, and what might be done to mitigate the risk. The risks are organized into three groups as follows:

GOAL RISKS

- Limited, ill-defined, or undefined goals
- Mis-application or 'off-label' use

DATA RISKS

- Bias inherent in the training dataset
- Irrelevant data in the training dataset
- Gaps or limitations in the training dataset
- Stale data in training datasets

RISKS WHEN USING TOOLS

- The unexplainable or "Black Box" aspect
- Inadequate performance measures
- Undefined accountability
- Fooling or hacking
- Loss of subject privacy
- Failure to Obtain, or Loss of, Trust

GOAL RISKS

RISK: Undefined, Ill-defined, or Limited Goals

(41) If AI tools are developed without an explicit expression of the task-specific goals the tool is designed to address, or the goals are incomplete or ill defined, the resulting tool may not achieve the intended goals when used and may even have unintended consequences or harm subjects. [8]

(42) Similarly, if the design and application of a tool does not support society's values governing the human activities that include the task for which the tool was developed, it may lead to loss of trust in the tool, if not harm to subjects.

SOURCES OF ILL-DEFINED GOAL RISKS

(43) The failure to explicitly identify and incorporate overarching values, such as the rule of law or basic human rights, during the design phase for a tool may produce a useless or even malicious tool. The tool's development process may 'find' a method or practice for its predictions that is inconsistent with basic values.[9] In the legal setting if the overarching goals and values of the rule of law are not addressed in the tool design and development phases, there is a risk that the tool will produce results that are counter to the rule of law, or that undermine it. For example, if a pretrial release tool results in over-incarceration or racial or wealth bias in pretrial incarceration, even if unintentional, it fails to achieve an overarching goal of being just.

(44) Lack of consensus among the design team about goals, especially if there are multiple, potentially conflicting goals, can also be a source of risk.[10] For example, in the field of risk assessment, is the goal of the tool to reduce incarceration, reduce recidivism, eliminate racial disparities, or further the fundamental rule of law principles of equality and fairness? A tool developed to achieve only one of these goals, or developed without the proper balance of multiple goals, may not be effective overall and may be unfair across groups or harmful to individual subjects.

(45) There are also risks from the use of a surrogate measure of whether a goal is achieved in the development and evaluation of a tool, which measure is subtly different from, and a less relevant than, the best measure of the goal. For example, a pretrial release tool that looks at the likelihood of subsequently being arrested or convicted of a crime (because datasets containing this data point are readily available) when the

original goal was to look at the likelihood of subsequently actually committing a crime, even if not arrested, for which there are no datasets available. In this example, the proxy measure will introduce bias inherent in the intervening activities of policing, arrest, and charging practices.

(46) Some AI developed tools have been found to have misinterpreted goals or come up with predictions that achieve the specified goal, but in a non-sensical or unexpected manner.[11] The result can be unintended, if not harmful, consequences. A simple example is a spam removal tool that achieves 100% spam elimination by moving all incoming email to the spam folder.[12] Another, horrific, example is a tool recommending purchases that suggests to those looking to buy cotton rope that they also consider buying a short wooden stool.[13]

MITIGATION OF ILL-DEFINED GOAL RISKS

(47) When the development of a tool is first conceptualized, the team should address the threshold question as to whether a technical, AI based, solution is even appropriate, given the project's goals.[14] In the legal setting, is it appropriate to develop a tool that makes a prediction regarding a particular person based on data about the behavior of other people the tool finds to be similarly situated? Simply stated, what many AI algorithms do is use data about groups to make predictions about an individual, an approach that is problematic in a legal setting that contemplates 'individual justice in individual cases'.

(48) At the beginning of the design process, explicitly identify and prioritize the goals the tool is intended to support before the training dataset is collected, before evaluation criteria are selected, before a tool is actually developed, and before the tool is applied in a particular setting.[15] Achievement of these goals should be explicit in the evaluation criteria or tests being used to assess the accuracy and validity of the tool.[16] Care must also be taken to not select a goal just because it is more readily measured, but which can unintentionally introduce bias or unfairness into the results.[17]

(49) From the beginning of the tool development process involve a full range of stakeholders, not just tool developers, in the articulation of goals. The stakeholders should include management, subject matter experts, and representatives of operators and subjects as well as data scientists and AI architects. [18] If the tool developers understand from this group what the overarching and fairness goals are, they can more effectively assess what data to collect and what models to build. In the justice system the overarching justice system goals and values which the tool must support should be explicitly identified and incorporated in the design. [19] For example, in developing tools for use in legal decision making, it is important to ensure that the values underlying the rule of law are incorporated into the development and evaluation of the tools. A significant aspect of this is to recognize that there are two dimensions to justice system goals and values – at the individual level and at the societal level.[20] Design and

evaluation of a tool should be at both levels, understanding that it may appear that they are not always in sync.

(50) Specifically, the following goals and values should be incorporated in the development and deployment of AI developed tools in a justice setting:

a. *Access to justice* – AI developed tools and their use should increase equal access to justice, in particular, for those traditionally underserved.

b. *Equal justice for all* – AI developed tools and their use should provide greater consistency and fairness of outcomes across subjects. The tool should also be impartial, minimizing the impact of irrelevant, inappropriate, or extraneous factors that could affect outcomes.

c. *Due process protections* – AI developed tools and their use should protect due process rights enumerated or inherent in the U.S. Constitution and the applicable state constitution.

d. *Procedural due process* – AI developed tools and their use should provide adequate and timely notice of proceedings, a hearing before an impartial judge, and an opportunity to be heard. The tool and process should also support the elements of respect, voice, neutrality, and trust.[21]

Again, achievement of these goals should be explicit not only in the design stage, but also in the evaluation criteria for testing and in validating the efficacy of a tool.

(51) Build and test several versions of a tool using different AI algorithms and compare the effectiveness and fairness of each tool relative to the overarching goals and values to identify the most just, as well as the most effective, tool.

(52) Where there are multiple goals, some of which may be conflicting, build different versions of the tool to address different goals (for example, recidivism and failure to appear), and provide the results for each goal when evaluating the tool. Alternatively, provide the results to a human team which applies 'weights' to each objective which weights are then used to make a prediction from a combination of the predictions of different tool versions.

(53) It may become possible to 'teach' an AI tool some aspects of context or understanding by including rules regarding context that support overarching goals during the development process. When mining a large training dataset, an AI developed tool converts statistical correlations into a form of implicit knowledge. Adding context (domain knowledge) as an explicit rule during a tool's development could 'speed up' a tool's learning as well as better support project goals.

(54) Check for consistency with overarching system values and unintended consequences when evaluating the overall effectiveness and fairness of the tool to accomplish what is intended.

(55) When using a tool, be transparent about the goals the tool is designed to support in order to allow operators and subjects to assess its applicability to their situation.

(56) Require that the tool prediction be only one factor considered by the operator and subject, alerting them to the need to consider the impact of other goals or the possibility of another option or recommendation.

(57) Don't rely solely on tool prediction – require human review and allow 'rebuttal' arguments or consideration of alternative decisions.

RISK: Goal Creep or "Off-Label" Use of a Tool

(58) If an AI developed tool is applied in a new setting where the goals are different, narrower than, or broader than, the goals for which the tool was originally built, it could result in unintended consequences or harm to subjects, as well as failure to achieve the goals of the new setting.

SOURCES OF GOAL CREEP OR "OFF-LABEL" USE RISKS

(59) A tool intended for a specific set of goals and objectives is applied without modification in setting with different goals or objectives. For example, a pretrial release tool whose development focused solely on likelihood of appearing at the next court date, but not recidivism, or that focused on specific crimes (such as violent crimes) rather than overall crime risk, is used as a release predictor where the goals are reduced recidivism or for different categories of crimes.

(60) Another form of 'goal creep' is when a new goal is superimposed on the use of a tool developed for other or fewer goals. For example, adding a requirement that use of a gun in a crime prevents pretrial release, where a release tool indicates the person is not a high risk.

MITIGATION OF GOAL CREEP OR "OFF-LABEL" USE RISKS

(61) Before applying a tool to a new setting, compare the goals of the new setting to the goals for the setting in which the tool was developed to see if they are compatible. If there are more or fewer goals, new, or conflicting goals, a new tool development process should begin, with an appropriate training dataset.

(62) In analyzing the appropriateness of the tool in the new setting, assess the validity and fairness of outcomes in the new setting relative to the goals of the new setting.

(63) If new goals are to be added, or a rebalancing of existing goals is contemplated, analyze the validity and fairness of the tool in the new setting to see if the results are acceptable or if a new tool should be built.

DATA RISKS

RISK: Bias Inherent in the Training Dataset

(64) The most significant, and insidious, source of risk is bias inherent in the training dataset used to develop or update a tool. The existence of bias in the training dataset can result in inappropriate, discriminatory, unequal, unjust, or unfair treatment of subjects to whom the tool is applied.[22]

(65) The term 'bias' has specific technical meanings in mathematics, statistics, and even in AI tool development. However, the use of the term bias here has to do with differential treatment of individuals or groups. Some forms of bias are legally prohibited, and AI tools could, if not developed with these laws in mind, result in predictions that are legally prohibited. In addition, in building tools developers may incorporate policies and practices, implicitly or explicitly, that may result in an unfair impact, on groups or individuals, or other forms of bias when used in the legal system.

(66) A simple example of training dataset bias is the use of historical data regarding failure to appear and recidivism in developing a tool to be used in supporting pretrial release decisions. The training datasets used to develop the tools generally consist of past decisions by judges releasing people who have been arrested. The implicit bias, even unconscious or unintentional, on the part of arresting agencies, prosecutors, and judges will be reflected in the training dataset. This systemic bias will most likely be detected and incorporated into an AI developed tool and, if so, the tool will exhibit a corresponding bias in making a release recommendation for an individual defendant.[23]

(67) In the family law context, there may be a bias in an AI developed tool suggesting parenting options that is based on a training dataset of prior patterns of parental sharing in cases. It will also reflect a bias of judges who tended to prefer particular parenting patterns in certain situations, which may or may not be based on empirically based studies of the efficacy of parenting plans, or which may preclude other viable options from appearing in the training dataset. Decisions in other cases may contain bias based on gender, socio-economic status, contemporary social norms, etc. [24] Again, any bias in the training dataset will likely be repeated in the AI tool predictions.

SOURCES OF BIAS RISKS

(68) Bias inherent in the human decisions that form the training dataset used to develop or update a tool.[25] Bias may be introduced in a process that is a precursor to the process being modelled. For example, criminal cases filed in a court are the result of policing, arrest, and charging processes that lead to the filing of a court case. Bias introduced in any of these earlier stages will be reflected in the dataset about subsequent stages.

(69) Training datasets may contain data elements that are unrelated to the tool goals or the predictions it is intended to make, which introduce bias. Extraneous information

contained in the training dataset may be seen by the AI analysis as relevant factors in making a prediction thereby introducing bias, even if unintended by the tool developer or the particular AI approach.

(70) If there is continuous updating of the tool by applying it to new data points created by the use of the tool, there will be feedback loop that perpetuates any bias in the original training dataset. Even if bias in the original training dataset is somehow neutralized, if the tool continuously learns with new data, there is a risk bias will be re-introduced, for example, if any new data, not generated by the tool, came from sources where there was inherent bias.

(71) Limitations in the perspectives and experiences of the tool developers subtly or unconsciously introducing "like me" bias. This can range from differences in values and acceptable social norms among the tool developers as compared to the population of subjects to which the tool will be applied or to differences in the educational level of developers and subjects that affect transparency and usability of the tool. [26]

MITIGATION OF BIAS RISKS

(72) The most effective approach to limiting bias risk is careful and knowledgeable selection, collection, and transformation of the data points in the training dataset. Careful judgment in the selection and representation of data may be more important than the selection of an AI approach -- paraphrasing the old saying, "bias in, bias out". One classic mistake would be using a particular dataset because it exists, without examining the dataset to see if the data elements are really relevant and applicable given the goals the tool is meant to support, or the types of predictions sought. Specific approaches to consider to avoid or mitigate bias risks include:

(73) Use training datasets that are more inclusive of the general population of subjects to whom the tool will be applied.[27]

(74) Build training datasets with more detail and greater granularity as to data points deemed more relevant to the prediction sought.

(75) Tool developers should be aware that AI can sometimes identify factors that are, in effect, surrogate measures for intentionally excluded factors like race or ethnicity. For example, zip code or a home address could be surrogate measures for race or ethnicity. To address the impact, one approach would be a statistical analysis of the correlation between certain variables to see if they could possibly bias tool predictions. Another approach is to run the tool development multiple times, once using all data elements and, in subsequent iterations excluding possible surrogate data elements to see the impact of the excluded data. A third approach would be to look for disparate impact by developing tools using training datasets for each 'protected' subgroup to assess the impact of the overall population diluting, obscuring, or overstating relevant factors for subgroups.

(76) Analyze the training dataset itself, before using it to build a tool, and 'clean' it to eliminate potential bias, for example, from the use of certain data codes that might introduce bias. A statistical analysis can be done to identify possible bias introduced by certain suspect variables, with possible adjustment to the tool to neutralize the bias.[28] An interesting approach here is to invert the analysis.[29] Use algorithms to identify bias in the decision made by people, inform those making the decision of the bias identified, and have the decision makers change their behavior to minimize bias in future decisions used to create new training datasets.

(77) Select measures for testing the efficacy and validity of a tool that reflect outcomes that will have minimal bias, or greater fairness.[30] Care must also be taken in selecting metrics when evaluating a tool. There are several types of bias inherent in data sets and use of different metrics can affect the apparent fairness of tool outcomes.[31]

(78) Use products or applications now offered by several companies that detect bias or unfairness in the training dataset or tool outcomes.[32]

(79) Apply the tool to new datasets and check for bias and unfairness.

(80) Use different AI algorithms to develop alternative tools and compare outcomes to identify the most effective and least biased version for use.

(81) Use different AI algorithms or different versions of AI algorithms to develop multiple tools and compare the outcomes to identify the most effective and least biased version for use. Consider combining the results of models built using different AI algorithms to generate a 'composite' score/prediction, an ensemble of the models. [33] By using a weighting formula, tool developers could balance the advantages and weaknesses of each algorithm to come up with a better overall predictive tool.

(82) Another approach to uncovering possible bias would be to conduct a "discrimination impact assessment" or "algorithmic impact analysis"[34] of the results of an AI developed tool to look for bias. The assessment would be similar to the "privacy impact assessment" (PIA) recommended for data privacy analysis and the "environmental impact report" (EIR) required for potentially environmentally sensitive policy decisions. Inherent in any such analysis is the difficult question of what benchmark(s) will be used when assessing the efficacy of an AI developed tool. Is it acceptable if the AI developed tool performs at least as well as humans, or will there be a higher standard for an AI developed tool? Neither humans nor AI developed tools can be expected to be error free, but what is the acceptable 'error rate'.

(83) Expand the pool of tool developers to reflect the broader cross-section of perspectives and experiences of prospective subjects and the population generally and train them to recognize bias in their thinking.[35]

(84) It is also important to provide transparency to subjects regarding possible bias in the training dataset used and its limitations.

(85) Finally, operators and subjects should not rely exclusively, or perfunctorily, on the tool. Rather, it should be used as a supplement or complement to human decision making.

RISKS: Irrelevant Data in the Training Dataset

(86) Training datasets may contain too much data, some of which is irrelevant to the goal and business objectives of the tool. An algorithm may find patterns in this extraneous information, identifying them as factors to be included in the prediction even though they may not make sense. For example, tools developed to read chest x-rays were found to base predictions in part on what x-ray machine was used to make the x-ray and the facility where the x-ray was taken.[36] Factors that are not included in a training dataset cannot affect the prediction offered by the tool. This is where the human ability to understand context and causation is especially relevant to counter an algorithm's inability to consider these perspectives.

SOURCES OF IRRELEVANT DATA RISKS

(87) Training datasets that contain data points whose relevance to the tool outcomes is not apparent. This may result from the use of an existing training dataset which has data relevant to the tool being developed, but also contains data used for developing other tools, or which just happened to be in the source of the training dataset. Care must be taken to collect data that has relevance to the prediction desired, and not just collect data, or use an existing dataset, because it is readily available.

(88) Training datasets may also contain data confounded by extraneous factors not included in the dataset and which may not be obvious. For example, studies have found decision by judges to be affected by factors such as time of day, hunger level, or concerns about unrelated events.[37]

MITIGATION OF IRRELEVANT DATA RISKS

(89) Before building a tool, analyze training datasets and only retain data elements that are relevant to the question being addressed and the project goals. Also, ensure that the coding values a data element can take are the values relevant to the prediction desired.

(90) Exclude data elements irrelevant to the intended predictions or inconsistent with the project goals or which are not currently used by human decision makers. Factors that are not included in the dataset from which the tool is built perforce cannot affect the prediction offered by the tool. This is where the human ability to understand context and causation is especially relevant to counter AI's inability to consider these perspectives. If the tool developers cannot see how a data element is relevant to the

goals and question to be answered, it probably should not be included in the training dataset.

RISK: Gaps or Limitations in the Training Dataset

(91) The content and nature of the training dataset can introduce many problems other than bias. AI analysis cannot identify relevant decision factors about which no data is present in the training dataset. The AI analysis is also constrained to the extent the training dataset reflects existing business practices and even the social, economic, and political relationships that underlie our institutions. 'Dirty' or 'noisy' datasets are also problematic. Finally, inconsistencies in the 'coding' the data in the training dataset may introduce irrelevant information.

SOURCES OF DATA GAP RISKS

(92) Training datasets that only include prior outcome patterns may not include all acceptable outcomes, and, obviously, do not include alternative, new, or innovative outcome patterns. Another problem is an AI tool's prediction for a situation where the dataset does not contain data about such situations. For example, a training dataset of past pretrial release decisions used to develop a pretrial release tool will only contain decisions by judges about defendants released, but no decisions or information for defendants that judges do not release at the pretrial stage because there is no data that can be included in the training dataset on the recidivism and appearance history for people who are not released.

(93) Training datasets that are too small or do not include a sufficient sample size of infrequent or 'fat tail' events will result in a tool less likely to successfully predict infrequent events or outlier outcomes. When predicting these rare events, AI developed tools can generate an unacceptable number of 'false positive' errors – recommending a wrong outcome – or fail to predict rare events at all. For example, the Extreme Vetting Initiative proposed by federal Immigration and Customs Enforcement (ICE) was intended to:

> "*evaluate whether a visa applicant: 1) would become 'positively contributing member of society;' 2) had the ability to 'contribute to the national interests;' and 3) 'intends to commit criminal or terrorist acts after entering the United States'.*"[38]

The third factor is an extremely infrequent occurrence[39] and it is not clear how large a training dataset or what characteristics need to be included in a training dataset to reliably predict anything about this type of factor. The risks are that the tool identifies people as terrorists who are not, or never predicts anyone to be a terrorist.

(94) Training datasets which include lots of missing data, null values, or 'dirty' data will weaken an AI analysis. 'Dirty data' is data that is inconsistently or inaccurately coded, data collected where there was a lack of standard identifiers and consistent

definitions, or data never audited to address these issues. The AI analysis may detect patterns that are attributable to missing or bad data as relevant to an outcome, resulting in the tool making mis-leading or even non-sensical predictions.

(95) Training datasets may include subjective data from which it is more problematic to detect useful patterns. For example, in the Extreme Vetting Initiative described above trying to predict whether a potential immigrant would become a "positively contributing member of society" requires definition of a "positive contributing member of society", something that appears to be quite subjective and difficult to 'code' in a dataset.

(96) Training datasets in courts that include data from legacy case management systems 'converted' to a newer data structure where data element definitions might not be the same or where there are data elements that require many-to-one combinations of old data elements or a one-to-many conversion of old data elements. This could be quite problematic, for example, in creating training datasets of family law factors and outcomes from which an AI analysis could build a useful tool.

MITIGATION OF DATA GAP RISKS

(97) As discussed above concerning bias risks, the most effective approach to limiting gap risks is careful selection of data points and the collection and transformation of data in the training dataset. Again, deliberate judgment in the selection and representation of data may be more important than the selection of an AI approach. Tool developers should also consider supplementing quantitative "hard" data with information gathered through qualitative methods about the intended subjects, so called "thick data"[40], to build a training dataset that better reflects the phenomenon to be modelled. Specific methods to consider regarding mitigating gap risks include:

(98) Use training datasets with a greater number of examples.[41]

(99) Use training datasets that are more inclusive of information about the factors expected to contribute to reliable predictions and more inclusive of the population and scenarios to which the tool will be applied.

(100) Use training datasets augmented to include all known viable outcomes that the tool should be able to predict. This should include outcomes that are considered viable that are not represented in the available training dataset, and for which examples need to be added. For example, use of home detention, an electronic bracelet, or day reporting center for pretrial release where the option has not previously been available in the jurisdiction.

(101) Use training datasets augmented to include more infrequent or 'fat tail' events, outliers, and niche cases, or use different approaches recommended for analyzing

datasets with these gaps[42], especially if the tool is intended to 'predict' infrequent or uncommon events.

(102) Evaluate, clean, and maintain consistency in the data definitions and codes used in the building the training dataset. An effective training dataset will reflect the input of people with deep domain knowledge about the system the data measure/reflects, and care in using data measures that are relevant to the problem to be solved.[43] It may also help to transform data to organize and contextualize it. For example, a data variable with three options might be more meaningful if it is transformed into three variables, each with a yes-no option.

(103) Missing and 'dirty' data, or data converted from legacy case management systems should be cleaned before it is used to build a tool.

(104) Less use of, or reliance on, training datasets containing more qualitative or subjective data whose meaning may vary, or depend on context, or on who coded the data.

(105) Capture more data than what is being collected now, both greater detail and detail more relevant to decision factors which could result in more successful tool predictions. Note that data in court case management systems should, and generally do, only contain data elements relevant to the court processing of a case. Training datasets needed to build a useful tool in court settings would generally require more, and different, data than this limited set. Significant data mining of court documents using data extraction algorithms may be necessary to build robust datasets for AI analysis. In addition, data not generally included in court documents may need to be collected from litigants. It is important to remember that the data relates to persons, that the decisions resulting from the use of a tool are about people and their problems. Augmenting the data in a court case management system should, therefore, be person-focused, not data-focused.

(106) Look for so called unknown unknowns and adjust gaps identified in the training dataset. This is obviously easier said than done.

(107) Provide transparency to operators and subjects regarding possible gaps, and their nature, in the training dataset used. Providing transparency to subjects about the training datasets and its limitations allows the subjects to assess the applicability of the AI tool's prediction to their particular situation.

(108) Finally, operators and subjects should not rely exclusively, or perfunctorily, on a tool. Rather, tool predictions should be used as to supplement or complement human decision making.

RISK: Data in Training dataset Becomes Stale

(109) Many AI algorithms learn to predict based on a static training dataset assembled when the tool was first being developed. The initial training dataset reflects the scenarios to be predicted as of a fixed point in time – when the data set was assembled. If the scenarios change over time, in composition or outcome patterns, the original dataset becomes less and less reflective of the real world and the predictions made by tools built using the original training dataset will become less reliable.

SOURCES OF STALE DATA RISKS

(110) AI algorithms that learn to predict from a training dataset that reflects past practices and experiences will generate predictions reflecting these past patterns. However, if patterns shift over time, the predictions will become less reliable and more 'dated'.[44] In addition, if patterns have changed over time in a dataset that goes back some period in time, predictions may overweight earlier patterns and underweight more current patterns. A very timely example is the effort of social media companies to keep up with hackers and others constantly shifting their algorithms and identities to promulgate mis-information.

(111) Training datasets that reflect patterns and practices which have become obsolete, changed in frequency of occurrence over time, or are slowly shifting will result in an AI tool whose predictions become less reliable. This is particularly problematic where the desired predictions are for events that might take place over a period of several years, for example, the likelihood of committing a crime during the next three years or the likelihood of an immigrant becoming a contributing member of society over a lifetime. The longer the time frame, the staler the data in a training dataset will become.

(112) If new outcomes are developed for scenarios, which outcomes are not contained in the training dataset, the tool can never predict the new outcome. If there are subsequent changes in the frequency of outcomes that are present in the training dataset, the tool will not accurately predict these shifts, as it is not aware of the changes.

(113) Changes in factors defining acceptable outcomes or addition of new acceptable outcomes will render a tool's predictions outdated. Also, if new criteria or rules are imposed governing acceptable outcomes, a tool's predictions may be wrong. For example, adding a criterion for pretrial release to include whether there was violence in the commission of the alleged crime will render a release tool ineffective and unusable.

MITIGATION OF STALE DATA RISKS

(114) When building a tool, consider weighting the data based on age, with older data having a lower weight.

(115) Identify the critical features which affect predictions and monitor their distribution in new scenarios to detect shifts which suggest the need to revise the tool.

(116) Review the scenarios where the tool made a wrong prediction to see if there are features that could be added to improve predictions and rebuild the tool.

(117) Periodically review current scenarios to which a tool would be applied to see if there are new acceptable outcomes, shifts in frequency of acceptable outcomes, or changes in the 'rules' governing the application. If found, revise the tool.

(118) Consider gathering new data as time goes forward to generate a more current, datasets that will better reflect current patterns and practices and revise the tool based on this new data.

(119) Consider building a tool with continuous learning ability.

(120) If new factors are added to determining predictions, rebuild the tool to include the new factors.

RISKS WHEN USING TOOLS

RISK: The Unexplainable or "Black Box" Aspect of Tools

(121) Many existing AI algorithms used to develop tools currently preclude tool developers, operators, or subjects from determining the basis of a prediction.[45] Not knowing how a specific prediction is made also prevents assessment of the source of apparent bias or other problems exposed when a tool is used. In a legal setting the inability to provide a transparent explanation of the prediction made may raise due process and other legal issues.[46] Currently it is difficult both to understand how a prediction was reached as to a particular scenario or person (referred to as 'local interpretability') or to understand whether predictions across scenarios are supporting over-arching goals (referred to as 'global interpretability').

SOURCES OF "BLACK BOX" RISKS

(122) The process through which AI generates a tool is often quite opaque such that humans cannot follow the trail of what features (data elements describing a scenario) a tool considered, which features, or combination of features, were most salient, how important a particular feature was, or more generally how the tool reached the prediction it offered. Unfortunately, more complex AI algorithms, such as deep reinforcement learning, although often more accurate, are less interpretable, than simpler algorithms. In addition, human intuition and prediction becomes weaker as the number of factors or dimensions increases, making it harder to for operators and subjects to intuitively accept a prediction where there are a large number of features considered by a tool.

(123) Opaqueness is exacerbated when the source and contents of the training dataset are unknown. Collecting and cleaning relevant data can be time consuming and expensive. Developers sometimes skip this step by using an existing dataset. If the source of the training dataset is not known, or its contents not examined before using, it is difficult to assess whether the source of any bias or other negative consequences is in the training dataset or the AI algorithm used to develop the tool.

(124) Finally, tool developer's interests in protecting intellectual property rights in the tool may result in an unwillingness to reveal the training dataset, the relevant features, how AI was used, or what AI algorithms were relied on in developing the tool. Again, this can prevent the building of trust necessary to the acceptance of the use of the tool.

MITIGATION OF "BLACK BOX" RISKS

(125) What are needed are "explainable AI" tools. Interpretability can be added when the tool is being built,[47] improved by choice of algorithm,[48] or applied after the tool is developed.[49]

(126) Develop some level of transparency or interpretability as to the inner workings of the tool to answer the question "why did the tool recommend X". [50] For some AI algorithms, it is possible to identify, at least at a high level, what features were considered, which features, or combination of features, were most salient, and how important a particular feature is to the prediction made.[51] Recently, a few technical approaches for providing an explanation have been made available, particularly for tools using deep learning algorithms.[52] Other approaches being explored include having the tool prepare a 'log' of its path to the prediction it makes, 'reverse engineering' a tool to determine how it made a prediction, and applying AI to a tool to try and reveal its 'thinking' processes.[53] Note that it is not necessary to be able to explain every detail of the AI tool's decision making process any more than it is necessary to explain every mechanical detail of the operation of a car to trust that it will operate safely and reliably.

(127) Open tools, and the training datasets upon which they are based, for inspection, analysis, and audit[54] to allow operators, subjects, and the public to understand the training datasets and AI algorithms used in developing the tool. Alternatively, a tool could be required to have an app allowing people to 'test' data to see how it works and check for bias. Several other related non-technical approaches include:

a) require the use of open source software to develop the tool;
b) require the use of an open/shared public training dataset in the development of a tool;
c) require commonly accepted or formally adopted standards to be used in the development of the tool;

d) require a tool to be analyzed or audited by an independent entity or regulatory agency similar to FDA review of new drugs; or
e) require a tool to be 'certified' by an oversight or regulatory entity.

(128) Another approach, directed more at the results than the internal processes, is to develop approaches that evaluate a tool's effectiveness and fairness relative to the defined goals of the tool. This may involve applying the tool to different datasets to identify possible bias, unintended consequences, or impacts. Another approach is to compare results from using different AI algorithms. Similarly, the impact of a particular factor can be examined by deleting or by changing the value of the factor, rerunning the AI analysis, and looking at the impact on predictions. Assessment of the impact of a tool is another alternative.[55]

(129) If the developer of the tool is unwilling to open the tool up for general inspection in order to protect intellectual property (IP) rights,[56] one option may be to allow 'in camera' investigation of the tool, as is commonly done in court hearings involving IP rights. The tool's source code might not need to be made available if validation studies and expert testimony about the tool are sufficient for the court to assess admissibility.[57] Another option would be to require a "discrimination impact assessment" or "algorithmic impact statement" about the tool to be publicly available.[58]

(130) Provide education and training to justice system operators who use a tool regarding the potential biases and fallacies of the tools so that their assistance is more informed and there is greater transparency.

(131) Notify the subject when an AI developed tool is being used to assist in making a legal decision affecting them or in helping them navigating the legal process. [59]

RISK: Inadequate Performance Measures

(132) Accuracy of some tools depends upon the tool's performance metrics for prediction. For example, one study used a publicly available facial recognition tool to match pictures of every current member of the House and Senate against a database of 25,000 publicly available arrest photos. Using the default match setting recommended by the tool provider, an 80% confidence rating, 28 members of Congress were falsely matched with mugshots.[60] The tool provider repeated the scenario using an 850,000 face dataset and a 99% confidence rating and got 0 matches.[61] [FN?] The metrics used, and the appropriate values of the metrics can significantly affect predictions, and public trust, in a tool.

SOURCES OF INADEQUATE PERFORMANCE MEASURE RISKS

(133) Choice of a performance metric that does not directly or adequately measure performance relative to the tool's goals.

(134) Setting a performance metric value so low such that the predictions result in too many false positives.

MITIGATION OF INADEQUATE PERFORMANCE MEASURE RISKS

(135) Careful selection of a performance metric that is most relevant to the tool's goals.

(136) Selection of a performance metric value that achieves the goals and business objectives set at the start of the tool building process.

RISK: Undefined Accountability

(137) While the use of an AI developed tool is normally expected to be successful, there will be scenarios where the impact or outcome is negative or harmful to someone. In such a scenario, the questions arise as to who is accountable for a harmful result? What liability does the developer of the tool have? What liability does the agency have for procuring the tool and using it? What liability does the operator of the tool have? What about the agency or public officials who use a tool to make decisions about people appearing before them, or who they serve? As with any new innovation, addressing accountability issues usually lags the deployment of a new tool. [62]

SOURCES OF ACCOUNTABILITY RISKS

(138) Lack of explicit policy establishing and distributing responsibility and accountability.

(139) Relying on the tool exclusively or without meaningful human review of the goals, objectives, or development of a tool, or the predictions made by a tool.

(140) Failure to assess a newly developed tool regarding potential harmful impacts. There needs to be an assessment of both potential negative impacts and any disparate negative impacts on legally cognizable or protected groups of subjects.

(141) Use of tools for purposes not contemplated in the original goals of the tool, so-called "off label" use.

MITIGATION OF ACCOUNTABILITY RISKS:

(142) The law or regulation authorizing use of the tool must state who is accountable, in particular, the obligations and responsibility of the developer who built a tool, the agency using the tool, the leadership in the agency, and each type of operator, when using the tool.

(143) AI developed tools should be used for decision support, not decision making. While this adds a possibly accountable person, the decision maker, it also reduces the risk of harm from unreviewed reliance on the tool prediction.

(144) Require that the results of the tool only be one factor considered by the operator and subject, alerting them to the need to consider factors not included in the tool.

(145) Subjects must be informed that they are subject to an AI developed tool, the goals of the tool, and the nature of the training dataset used to develop the tool.

(146) Train operators on proper use of the tool, including understanding its underlying objectives, weaknesses, and limitations, and the nature of its predictions.

RISK: Fooling or Hacking a Tool

(147) Recent experiences suggest that variances in inputs, whether intentionally, unintentionally, or mischievously introduced, can cause AI developed tools to make unreliable outcome predictions.[63] The predictions made in response to the altered scenario range from harmless and humorous to potentially detrimental.

SOURCES OF HACKING RISKS:

(148) If a hacker has knowledge of what factors are relevant to generating particular outcome, the hacker can intentionally introduce incorrect, bad, even random data into a scenario to which the tool will be applied, to generate an erroneous outcome. One example is Microsoft's experience with the TAY chatbot which, in less than a day, hackers had 'trained' to exhibit an appalling personality, spouting crude comments and repeating racist statements and conspiracy theories.[64]

(149) Significantly, changes to input data values that are imperceptible or seemingly irrelevant to humans can cause a tool to make erroneous outcome predictions. For example, putting a small day-glow sticker in the corner of an image of a banana caused a classifier tool to label it a toaster.[65] In another example, a very slightly altered picture of a cat resulted in the classifier predicting it was guacamole.[66] An example with a more serious consequence is the presence of graffiti on a stop sign that could cause a driverless car not to stop when it should.

MITIGATION OF HACKING RISKS:

(150) Seek ways to increase the explainability of a tool to improve the ability to detect hacking. Another approach would be to test the tool to assess sensitivity to potentially misleading data either based on review of predictions that appear to be unusual or outliers or to highlight which data are most critical to a prediction and thus where hacking could be most damaging.

(151) Establish procedures to maintain the integrity of the tool use by validating the input data to protect against unauthorized introduction of, or changes to, the data.

(152) Provide for human review of each predicted outcome to detect anomalies or outcomes that seem inappropriate or improbable.

RISK: Loss of Privacy of Subjects

(153) Loss of privacy can occur both with regard to individuals whose data is in the training dataset and in the application of tools to subjects.

SOURCES OF PRIVACY RISKS

(154) Training datasets can contain information on a large number of individuals. In some instances the dataset includes personal identifying information (PII) about individuals obtained without their knowledge or consent.[67] Training datasets with greater detail and granularity about each individual increases the amount of information that may be considered private. For large datasets obtaining permission from each individual whose PII is in a training dataset may not be practical and may be perceived as slowing the development of AI based tools.

(155) Training datasets with more data points about each scenario and a greater volume of scenarios can result in identification of more subtle or non-obvious information about individuals.[68] Some AI algorithms predict the status or condition of an individual based on combinations of data points in the training dataset that are not otherwise known or obvious, perhaps even by the subject, or which are a 'false positive', that is, untrue. Exposure or use of the prediction may result on the loss of a subject's privacy.

MITIGATION OF PRIVACY RISKS

(156) Obtain informed consent from individuals whose data will be included in the training dataset. The consent should be based on an explanation of the data collected and the purposes for which it will be used. Individuals should also be provided meaningful choices regarding the collection, use, and sharing of their personal information.[69]

(157) Anonymize or encrypt data in training datasets or limit access to person-specific PII in training datasets during the development and during review or audit activities related to the adoption of the tool.[70]

(158) Adopt policies and protocols to limit access by operators to information derived by the AI tool that is not directly relevant to, or part of, the prediction for which the tool was developed.

(159) Share tool predictions in a manner that respects personal privacy and provides options for appropriately limiting the disclosure of certain types of information.

RISK: Failure to Obtain, or Loss of, Trust in Use of a Tool

(160) It is important to understand the wide range of trust individuals exhibit regarding predictions offered by a computer. Some people may initially not trust a prediction from a computer. Others do not trust human 'experts'. Consequently, AI tools should be built and used in a manner that recognizes these divergent perspectives.

SOURCES OF LACK OF TRUST RISKS

(161) Some people are uncomfortable with any prediction if it is offered by a computer algorithm. Some will not trust a tool if they perceive, rightly or wrongly, that it does not

incorporate factors they deem relevant to a prediction. For some people the fact that a tool uses statistics and is not deterministic will reduce their trust.

(162) Another group of people do not, or no longer, fully trust human decision makers, in particular, experts, and may be more trusting of a decision offered by a computer. Some may even assume a tool prediction is superior to a human prediction.

(163) The ambiguity of the source of the training dataset and whether there was appropriate consent to the use of the training dataset, either by the subjects of the dataset, or by the entity aggregating the dataset, can reduce or preclude trust in the tool predictions by operators and subjects.

(164) Unwillingness to provide information about a tool in order to protect intellectual property rights can prevent acceptance of the use of the tool.

(165) A recent public survey[71] found several themes regarding people's reluctance to accept decisions made by a computer. The themes included: a) they violated privacy, b) they were unfair, c) they remove the human element from important decisions, and d) humans are complex, and these systems are incapable of capturing nuance.

MITIGATION OF LACK OF TRUST RISKS

(166) Provide greater transparency to subjects and the public about the decision to build a tool, its goals and business objectives, the training dataset used to build a tool, the algorithms used, and the performance metric and values relied on to justify the use of the tool.

(167) If the tool provider is relying on intellectual property concerns to decline transparency of the tool itself, provide analysis of the outcomes, audit, algorithmic impact analysis or similar evaluation to demonstrate the effectiveness and fairness of the tool.

(168) Use human decisions and tool predictions in parallel until trust in use of the tool is established, or not. Once accepted as effective, continue to use human decision-making in conjunction with tool predictions. Care must be taken that over time operators do not become complacent about use of the tool and stop paying attention to the efficacy of its predictions.

HUMAN DECISION MAKING IN COMPARISON

(169) The discussion above has identified benefits and risks associated with use of AI developed tools. To keep this in perspective, it is appropriate to also examine benefits and risks associated with human decision-making. Comparing the advantages and disadvantages of human and AI approaches can suggest complementary approaches for achieving better overall results.

(170) Humans generally approach decisions with a quite significant store of information – prior knowledge – in their 'training dataset'. Not only do they have the information from personal experiences during their lifetime, they can also access information accumulated by humanity over time, including what is coded in our DNA through evolution. Although humans cannot know everything all other humans know, they know information is there to be accessed and how to access it. They also have a lifetime of building cognitive models (not always perfect) of how the world works and learning strategies to apply to new situations.

(171) Humans also suffer from gaps in knowledge. Decisions made by humans may not be consistent with empirical knowledge about the matter to be decided. Humans may not know something, or may choose to ignore it. They may also rely on inaccurate or incomplete information, sometimes unknowingly.

(172) Humans have also developed a set of decision-making approaches different in kind from AI approaches. Humans engage in deductive and inductive reasoning. Humans can use insight, intuition, heuristics, inference, and short-cuts when making decisions. Humans can also assess the context of information and can often deduce the causal relationships underlying activity. They can also think about their thinking processes – the context of their thinking. In addition, humans can use these approaches sequentially or in parallel.

(173) Humans can apply understanding, knowledge, and skills learned in one setting to another setting. This is due, in large part, to humans' ability to discern relationships and underlying causality in how a process works, and apply that to other, similar, processes.

(174) Humans can also be creative, thinking of new approaches, practices, or outcomes that have not existed before, or were previously unknown to them. Humans can also learn from their mistakes, and the mistakes of others, often the source of creative thinking. And these new approaches are generally consistent with cognitive models of how the world works, based on causal relationships, and can incorporate accepted social norms and values regarding acceptable behaviors.

(175) Humans can also 'fill in the blanks' or use judgment in ambiguous situations, in particular, taking into account context. Some legal decisions involve choices of outcomes whose effectiveness may vary depending on the people and circumstances involved. For example, the most appropriate parenting plan for a couple with young children will be very different from a parenting plan involving children who are in middle school or high school. Seldom does the law specify outcomes based on this level of detail, rather, it directs that the parenting plan should be "in the best interests of the child". Humans have a history of making decisions in these types of settings that can be used when making subsequent decisions.

(176) At the same time, human decision-making can be clouded by passions, emotions, apprehensions, anxiety, preferences, habits, and bias. There are innumerable examples of various forms of bias in legal decisions made by humans, in individual cases and collectively across cases. Humans sometimes rely on mis-information, confirmation bias, or are affected by extraneous factors such as fear, hunger, fatigue, etc. Any of these can degrade the quality and accuracy of human decisions.

(177) Human decision-making can significantly degrade where there is an over-abundance of information, whether in terms of volume, dimensions, complexity, or speed of delivery. The capacity of humans to reason where there is a high number of variables is limited, causing humans to resort to heuristics and short cuts that can sometimes be ineffective or suboptimal.

(178) Human decision-making can also suffer from undefined or conflicting goals. Incentives may also conflict with, or be inconsistent with, stated goals and values, sometimes because of old habits or bias, consciously or implicitly.

(179) Finally, in reaching decisions, humans are not always able to fully explain their decisions. Although judges generally provide an explanation of how they reached their decision in written statements of decision or opinions, not every step or element of their decision making is always provided. Nor can it always be described.

(180) While there are plenty of real and potential problems with the predictions of an AI developed tool, there can also be plenty of real and potential problems with human decision-making. One objective when considering deployment of AI developed tools would be to assess when to use humans and when to use AI developed tools in concert to reach better decisions, and when one should rely mostly on one or the other.[72]

CONCLUSION

(181) AI developed tools can improve navigation of court processes and enhance legal decision-making. However, their development and use are not without risk. The risks identified here are not insurmountable; they can be mitigated with careful planning and mindful development of tools. The problems arising from the use of AI to develop tools noted here are not new. New technology has brought them to the forefront. But it is not just the technology that needs to be better understood. Using AI in the justice system must also recognize that the justice system is a network of social values, policies, rules, understandings, and relationships built by humans, not technology, to guide and constrain behavior of humans. Use of AI developed tools should be used to enhance human intelligence and decision making, not replace it.

(182) The use of AI to build tools used in courts requires us to be more precise in stating policy choices and to be more explicit and transparent about the goals these policies are intended to further. It is also clear that the time when risk mitigation

measures are most critical and effective is when a tool is being developed. The best tools will come from use of appropriate and well-developed training datasets, using multiple AI approaches to developing a tool, and an analysis of the impact of various tools on different groups and circumstances. This will require not only AI technologists, but data scientists who understand the datasets, ethnologists who understand the dynamics of the human relationships involved, subject matter experts, and representatives of operators and subjects. Transparency regarding all of these steps will significantly contribute to the development of trust needed for effective use of AI developed tools used in courts.

ENDNOTES

[1] The Legal Services Corporation (LSC) is developing a design for self-represented litigant (SRL) web portals in each state with very similar vision, see REPORT OF THE SUMMIT ON THE USE OF TECHNOLOGY TO EXPAND ACCESS TO JUSTICE, December 2013.

[2] For good discussions of preparatory steps in building a tool, see Cassie Kozyrkov, "Getting started with AI? Start here! Everything you need to know to dive into your project", HACKERNOON, Oct 19, 2018, and Rahul Agarwal, "Six Important Steps to Build a Machine Learning System, A field guide to thinking about ML projects", September 7, 2019. For a discussion of involving senior management at this stage, see Mala Anand, "Want Responsible AI? Think Business Outcomes", Knowledge@Wharton, July 17, 2019.

[3] For a discussion about the importance of performance metrics and a description of basic metrics, see Cassie Kozyrkov, "Getting started with AI? Start here! Everything you need to know to dive into your project", HACKERNOON, Oct 19, 2018, and William , Koehrsen, "Beyond Accuracy: Precision and Recall, Choosing the Right Metrics for Classification Tasks", March 3, 2018.

[4] For further discussion of the tool building process, see Jason Brownlee, "Applied Machine Learning Process", February 12, 2014, and Divya Poddar, "Machine Learning: A Beginner's Guide", UPX Academy, December 5, 2016.

[5] For discussions of issues regarding collecting effective training datasets, *see*: Pete Warden, "Why you need to improve your training data, and how to do it", May 28, 2018; Dipanjan Sarkar, "Understanding Feature Engineering (Part 1) — Continuous Numeric Data", January 4, 2018; Pradeep Menon, "An Executive Primer to Deep Learning", Feb 20, 2018; Jason Brownlee, "How to Prepare Data For Machine Learning", December 25, 2013 in Machine Learning Process; and Jason Tashea, "Want to improve AI for law? Let's talk about public data and collaboration", American Bar Association Journal, May 22, 2018.

[6] See "AlphaGo is the first computer program to defeat a professional human Go player, the first to defeat a Go world champion, and is arguably the strongest Go player in history".

[7] Kevin Kelly, "The AI Cargo Cult: The Myth of a Superhuman AI", WIRED, April 27, 20117.

[8] For a debate about incorporating overarching values and objectives in AI tools, see Ben Pace, "Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More", Less Wrong, Oct 3, 2019.

[9] Johannes Heidecke, "Risks of Artificial Intelligence", Thinking Wires, published July 5, 2017, last edit July 30, 2017.

[10] For an example of tests of facial recognition where there were issues regarding the lack of clarity of the goals supported by the tool, see Prof. Pete Fussey, and Dr. Daragh Murray, "Independent Report

on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology", The Human Rights, Big Data and Technology Project, July 2019.

[11] See examples described in Aaron M. Bornstein, "Is Artificial Intelligence Permanently Inscrutable?", Nautilus Magazine, Nov 11, 2018.

[12] Links to lists of other examples can be found in Victoria Krakovna, "Specification gaming examples in AI", April 2, 2018.

[13] Ben Gilburt, "AI — The control problem", May 24, 2018, at: https://towardsdatascience.com/ai-the-control-problem-c82bb485bc54.

[14] See Google PAIR Initiative, "People + AI Guidebook, User Needs + Defining Success", at: https://pair.withgoogle.com/chapter/user-needs/.

[15] For a discussion of the process, see Ingedata, "7 Fundamental Tips To Build Successful AI", September 17, 2018. For examples of how this can be done, see European Commission High-Level Expert Group on Artificial Intelligence, ETHICS GUIDELINES FOR TRUSTWORTHY AI, and IEEE, ETHICALLY ALIGNED DESIGN, A VISION FOR PRIORITIZING HUMAN WELL-BEING WITH AUTONOMOUS AND INTELLIGENT SYSTEMS, First Edition.

[16] Andrew D. Selbst, danah boyd, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi, "Fairness and Abstraction in Sociotechnical Systems", ACM Conference on Fairness, Accountability, and Transparency (FAT*), August 23, 2018, at: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3265913_code1328346.pdf?abstractid=3265913&mirid=1. For a discussion of optimization to achieve over-arching goals, see David Weinberger, "Optimization over Explanation", January 28, 2018.

[17] Karen Hao, "This Is How A.I. Bias Really Happens — and Why It's so Hard to Fix", MIT Technology Review, Feb 4, 2019.

[18] Anand Rao and Ilana Golbin, "What is fair when it comes to AI bias? It's not the algorithm behaving badly, but how we define fairness that determines an artificial intelligence system's impact", April 12, 2019.

[19] For a discussion of inserting values into the process, see Jesus Rodriguez, "Teaching AI Human Values, OpenAI Believes that the Path to Safe AI Requires Social Sciences", February 21, 2019, and the OpenAI report Geoffrey Irving and Amanda Askell, "AI Safety Needs Social Scientists", Feb. 19, 2019.

[20] Greg Berman, "Our 2 Kinds of Criminal Justice, and How to Reconcile Them. We need both micro justice and macro justice. But they aren't always in sync", Governing, June 7, 2018.

[21] For a discussion of the meaning of these elements, *see* Judicial Council of California, PROCEDURAL FAIRNESS IN THE CALIFORNIA COURTS, 2007.

[22] See, for example, Drew Harwell, "Amazon Facial-ID Software Used by Police Falls Short on Accuracy and Bias, Research Finds", Washington Post, Jan 28, 2019, and Roberto Iriondo, "Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women", Oct 11, 2018.

[23] This has been the subject of considerable discussion. See Karen Hao and Jonathan Stray, "Can you make AI fairer than a judge? Play our courtroom algorithm game" MIT Technology Review, Oct 17, 2019, Sarah Picard, Matt Watkins, Michael Rempel, and Ashmini Kerodal "Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness", Center for Court Innovation, July 1, 2019, Aaron M. Bornstein, "Are Algorithms Building the New Infrastructure of Racism? How we use big data can reinforce our worst biases — or help fix them", Nautilus, Feb 21, 2019, Parul Pandey, "Is your Machine Learning Model Biased? How to measure your model's fairness and decide on the best fairness metrics", February 6, 2019, Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julia Ciccolini, "Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment", Criminal Justice and Behavior Journal, November 23, 2018, and Rachel Cortland, "The Bias Detectives", Nature, vol. 558, p. 357, June 20, 2018.

[24] *See* Diana Yates, press release of study: "Judges as susceptible to gender bias as laypeople – and sometimes more so", April 19, 2018 8:30 AM, at: https://news.illinois.edu/view/6367/640610. The

study described is: "Expertise fails to attenuate gendered biases in judicial decision-making" by Andrea Miller, available at http://journals.sagepub.com/doi/full/10.1177/1948550617741181. For examples of other sources of bias, see Daniel L. Chen, "Machine Learning and the Rule of Law", January 6, 2019, and Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous factors in judicial decisions", April 26, 2011, PNAS, vol. 108, no. 17, pp 6889-6892.

[25] For a quick summary of types and sources of bias, see Elizabeth Wallace, "9 Common Mistakes That Lead To Data Bias", May 3, 2019.

[26] *See* Sarah Myers West, Meredith Whittaker, and Kate Crawford, "Discriminating Systems, Gender, Race, and Power in AI", AI Now Institute, April 2019, AI Now Institute, THE AI NOW REPORT, THE SOCIAL AND ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN THE NEAR-TERM, July 7th, 2016, Jack Clark, "Artificial Intelligence Has a 'Sea of Dudes' Problem", June 23, 2016, and Ziv Epstein, Blakeley H. Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan, "Closing the AI Knowledge Gap", March 20, 2018.

[27] Jonathan Vanian, "Artificial intelligence can imitate and enhance human decision-making -- and amplify human prejudices. Can Big Tech tackle A.I.'s discrimination problem?", Fortune, June 25, 2018, this article originally appeared in the July 1, 2018 issue.

[28] For en example of how to develop a classification algorithm achieving both group fairness and individual fairness, see Richard Zemel, Yu (Ledell) Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork, "Learning Faire Representations", C. Proc. 30th Int. Conf. Int. Conf. Mach. Learn. 28, III-325–III-333 (2013).

[29] Daniel L. Chen, "Machine Learning and the Rule of Law", January 10, 2019, forthcoming in COMPUTATIONAL ANALYSIS OF LAW, Santa Fe Institute Press, ed. M. Livermore and D. Rockmore.

[30] Michael Veale and Reuben Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data", December 2017, Big Data and Society, at: https://journals.sagepub.com/doi/pdf/10.1177/2053951717743530 and Open Data Science (ODSC), "Machine Learning Model Fairness in Practice ", October 10, 2019, at: https://medium.com/swlh/machine-learning-model-fairness-in-practice-bdebeaa76ee8

[31] Anand Rao and Ilana Golbin, "What is fair when it comes to AI bias?", Strategy + Business, Tech and Innovation, April 12, 2019, at: https://www.strategy-business.com/article/What-is-fair-when-it-comes-to-AI-bias?gko=827c0&mod=article_inline and Parul Pandey, "Is your Machine Learning Model Biased? How to measure your model's fairness and decide on the best fairness metrics", February 6, 2019 and references cited in the article.

[32] For example, see Jeremy Kahn, "Accenture Unveils Tool to Help Companies Insure Their AI Is Fair", Bloomberg News, June 13, 2018, 7:00 AM GMT and David Weinberger, "Playing with AI Fairness, Google's new machine learning diagnostic tool lets users try on five different types of fairness", Google – People+AI Research (PAIR).

[33] *See* James Le "12 Useful Things to Know about Machine Learning", January 26, 2018.

[34] See, for example, Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker , ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY, AI Now Institute, April 2018, and FAT/ML, PRINCIPLES FOR ACCOUNTABLE ALGORITHMS AND A SOCIAL IMPACT STATEMENT FOR ALGORITHMS.

[35] See Yael Eisenstat, "The Real Reason Tech Struggles With Algorithmic Bias, Opinion: Humans train the machine-learning and AI systems at Facebook, Google, and Twitter to filter out bias. The problem: they don't know what they're looking for", WIRED, 02.12.2019 09:00 AM, at: https://www.wired.com/story/the-real-reason-tech-struggles-with-algorithmic-bias/.

[36] John Zech, "What are radiological deep learning models actually learning?", July 8, 2018, at: https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98.

[37] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous factors in judicial decisions", PNAS, vol. 108, no. 17, pp 6889-6892, April 26, 2011.
[38] Faiza Patel, "Extreme Vetting by Algorithm", Just Security, Monday, November 20, 2017.
[39] There have only been 19 confirmed terrorists who highjacked flights in the US between 2000 and 2017, see list of aircraft hijackings at https://en.wikipedia.org/wiki/List_of_aircraft_hijackings#2000s.
[40] Tricia Wang, "Why Big Data Needs Thick Data", January 20, 2016.
[41] See Irene Chen, Fredrik D. Johansson, and David Sontag, "Why Is My Classifier Discriminatory?", Advances in Neural Information Processing Systems 31, pages 3543--3554. Dec. 2018, from 32nd Conference on Neural Information Processing Systems (NeurIPS 2018).
[42] *For example,* Jason Brownlee, "8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset", August 19, 2015.
[43] Jason Brownlee, "How to Prepare Data For Machine Learning", December 25, 2013 in Machine Learning Process.
[44] See Alexandre Gonfalonieri, "Why Machine Learning Models Degrade In Production", July 25, 2019.
[45] For a discussion of the nature of this risk, see Aaron M. Bornstein, "Is Artificial Intelligence Permanently Inscrutable?", Nautilus Magazine, Nov 11, 2018.
[46] See AI Now Institute, LITIGATING ALGORITHMS: CHALLENGING GOVERNMENT USE OF ALGORITHMIC DECISION SYSTEMS, September 2018, at: https://ainowinstitute.org/litigatingalgorithms.pdf.
[47] Bahador Khaleghi, "The How of Explainable AI: Pre-modelling Explainability", July 31, 2019.
[48]Bahador Khaleghi, "The How of Explainable AI: Explainable Modelling", July 31, 2019.
[49]Bahador Khaleghi, "The How of Explainable AI: Post-modelling Explainability", July 31, 2019.
[50] For a discussion of the interpretation of the results of an AI developed tools, see Daniel S. Weld and Gagan Bansal, "The Challenge of Crafting Intelligible Intelligence", Communications of the ACM, vol. 62, no. 6, p. 70, June 2019, Lars Hulstaert, "Interpreting machine learning models", February 20, 2018, and Dipanjan (DJ) Sarkar, "Explainable Artificial Intelligence (Part 1) — The Importance of Human Interpretable Machine Learning", May 24, 2018.
[51] *See, for example*, Parul Pandey, "Interpretable Machine Learning, Extracting human understandable insights from any Machine Learning model", April 2, 2019.
[52] For discussion of a tools that can assist with this, see Ann R. Thryft, "Not a Lot of Debiasing, Auditing Tools Yet", EE Times, May 5, 2019.
[53] *See* The Economist, "The Behavioural Ecology of Machines; To Understand Digital Advertising, Study Its Algorithms; a Skinner Box for Software", March 22, 2018, The Economist, "For Artificial Intelligence to Thrive, It Must Explain Itself", Feb 15th 2018, Cliff Kuang, "Can A.I. Be Taught to Explain Itself?", New York Times Magazine, November 21, 2017, and Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford, AI NOW 2017 REPORT.
[54] For a discussion about audit challenges see DJ Pangburn, "How to Lift the Veil Off Hidden Algorithms", Fast Company, Jan 28, 2019.
[55] See Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood, "Accountability of AI Under the Law: The Role of Explanation", ver. 2 at: https://arxiv.org/pdf/1711.01134v2.pdf.
[56] *But see* Jason Tashea, "Trade secret privilege is bad for criminal justice", ABA Journal, July 30, 2019, 6:30 am CDT; and Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System", 70 Stan. L. Rev. 1343 (2018).
[57] See example of access to information about a tool in Jason Tashea, "Defense lawyers want to peek behind the curtain of probabilistic genotyping", American Bar Association Journal, December 2017.
[58] See discussion in paragraph (82) above).

[59] For a discussion of debate about whether the new European Union General Data Regulation creates a 'right to explanation', *see* Andrew D. Selbst and Julia Powles, "Meaningful Information and the Right to Explanation", November 27, 2017, International Data Privacy Law, vol. 7, no. 4, pp 233-242 (2017).

[60] Jacob Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots", ACLU of Northern California, July 26, 2018.

[61] Monica Nickelsburg, "Amazon challenges ACLU study on facial recognition tech and police", GeekWire, July 27, 2018 at 1:26 pm.

[62] For a discussion of accountability issues of AI tools, *see* Cedric Villani, FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE, TOWARDS A FRENCH AND EUROPEAN STRATEGY, A parliamentary mission from 8th September 2017 to 8th March 2018, pp. 113-125.

[63] See examples cited in Melanie Mitchell, "Artificial Intelligence Hits the Barrier of Meaning", opinion, New York Times, Nov. 5, 2018.

[64] Michael Li, "Addressing the Biases Plaguing Algorithms", Harvard Business Review, May 13, 2019.

[65] M. Mitchell Waldrop, "What are the limits of deep learning?", originally published at PNAS on January 22, 2019.

[66] Jonathan Zittrain, "The Hidden Costs of Automated Thinking", he New Yorker, July 23, 2019.

[67] For example, a dataset of faces of individuals gathered without the consent of the individuals, see Madhumita Murgia, "Microsoft Quietly Deletes Largest Public Face Recognition Data Set, Stanford and Duke universities also remove facial recognition data", Financial Times, June 6, 2019.

[68] *For example, see* Matt Fredrikson, Somesh Jha, Thomas Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures"on identifying an individual's identity from a dataset.

[69] See Helen Nissenbaum's suggestions discussed in Scott Berinato, "Stop Thinking About Consent: It Isn't Possible and It Isn't Right", Harvard Business Review, September 24, 2018.

[70] Another option might be an approach referred to as 'differential privacy' when sharing data. See Tyler Elliot Bettilyon, "Why 'Anonymized Data' Isn't So Anonymous, Cleaning data of 'personally identifying information' is harder than you might think", Apr 24, 2019, and Tianqing Zhu, "Explainer: what is differential privacy and how can it protect your data?", The Conversation, March 18, 2018.

[71] Pew Research Center, "Public Attitudes Towards Computer Algorithms", Nov. 16, 2018.

[72] See Dawson Eliasen , "Data, Algorithms, and Humans Making sense of the Information Age", May 25, 2019.