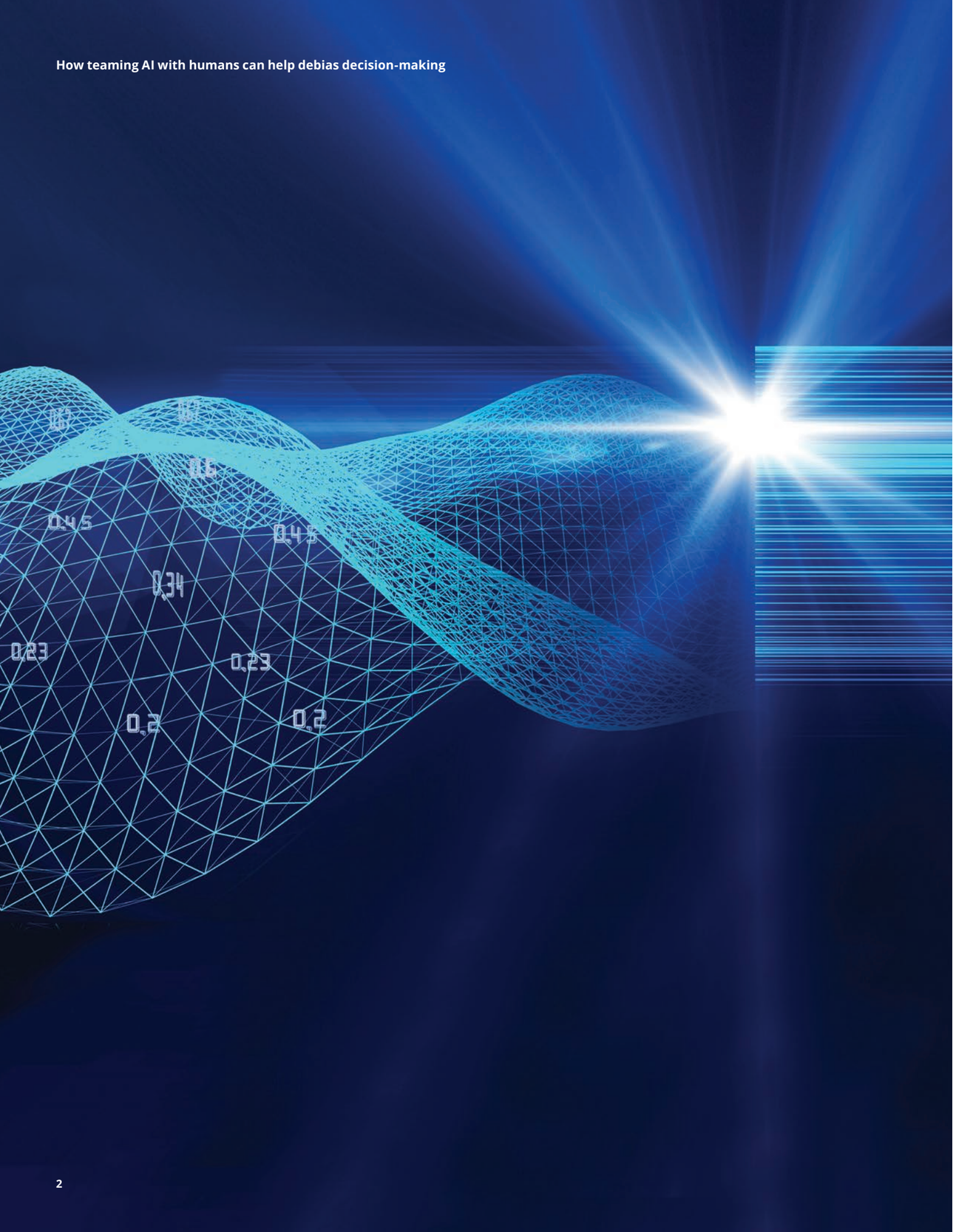
Deloitte.

# How teaming Artificial Intelligence (AI) with humans can help debias decision-making

By Tasha Austin, Joe Mariani, Devon Dickau,
Pankaj Kamleshkumar Kishnani, and Thirumalai Kannan

A publication from the Deloitte AI Institute for Government
July 2023








# At its core, government is a form of collective decision making.

Governments frequently use institutions and structures to prioritize how they spend money, decide on national defense, and seemingly mundane tasks like where to locate bus stops. So **improving decision making is synonymous with improving government.**

Enter, artificial intelligence (AI). AI has incredible power to find patterns in large amounts of data to help identify conclusions that human decision makers may not have been able to identify. Tapping into that power, governments have used AI to help allocate grants, prioritize health and fire inspections, detect fraud, prevent crime, and personalize services. However, AI may have programmed biases that systematically produce outcomes that may be considered to be unfair to one person or group. The challenges of biased AI algorithms are becoming well known. **From potential flawed facial recognition to potential biased bail decisions, having an over-reliance on AI may create significant challenges for government organizations.**

But hidden in those potential biases may be a path forward to even more equitable government where all people have the opportunity to thrive. **The limitations of AI and human decision making are the inverse of each other.** Where humans struggle with large volumes of data, precision, and consistency, AI can excel. Similarly, AI may struggle in adapting to context or understanding human values, things many humans do naturally almost without thought.

**Pairing human knowledge and experiences with AI capability** may allow governments the ability to tackle complex decisions, with greater confidence in the accuracy—and equity—of its conclusions. AI may help augment human capabilities by analyzing voluminous datasets and providing the ability to identify unconscious inconsistencies or potential biases in human judgments.

Managing human-machine teaming may be a challenge for an organization, but with effective governance, well-chosen data, and diversified talent, government organizations may use AI to improve the quality of their decisions.

And better, more equitable decisions mean better, more equitable government services.



How teaming AI with humans can help debias decision-making

# AI teaming in decision making

AI and human judgment may be perfect partners for each other. Human judgment may be wise and sensitive to context but has limitations. AI is very powerful but will only go where its programming directs it. What one lacks, the other may provide.

## Humans need data, consistency, precision

In his nearly 40 years with the Central Intelligence Agency, Richards Heuer analyzed some of the most important—and well-funded—decisions that human judgment could make. He found that not only was **human judgment inherently poor at dealing with the uncertainties of important decisions, but that merely being aware of potential sources of bias was not enough.**[1] For example, being too hot or too cold could influence decisions, but Heuer found that even knowing that fact did not help. Also, knowing that individuals have a predilection to discount information that does not fit initial assumptions does not improve the quality of decisions.

In other words, the **greatest strength of human judgment may also be its greatest limitation.** Individuals are able to sense context and adjust their thinking accordingly, something that machines cannot do. Yet, individuals are also susceptible to factors beyond their control, factors that individuals would not want influencing the decision. Those factors may range from the style of dress of a job candidate to the order in which individuals receive information. As Heuer's research points out, **humans are often not aware of how these factors may be influencing their decisions.** Compounding the issue, unconscious bias in human decision-making may further disadvantage groups of people who may have already been systematically marginalized.

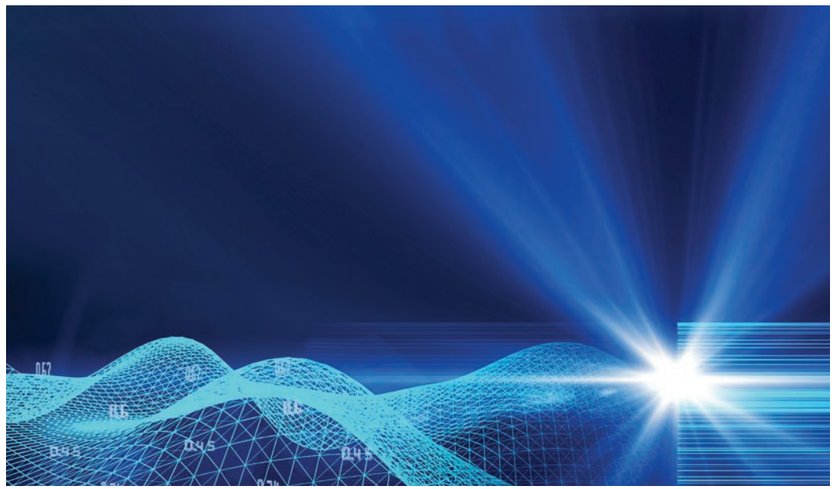
However, Heuer did find that tools designed to engage an analyst's higher-level analytical thinking could improve the quality and consistency of judgments.[2] This is precisely where AI can help. **AI's strength is the precision and consistency that human judgment lacks.**[3] So using AI as a tool to engage higher-level analytical thinking may help improve human decision making in areas where it may be weakest. AI may powerfully analyze data without being distracted by any factors other than what it was programmed to do.

## Machines need context and human values

However, government agencies should understand that **for all its power, AI is only a tool. It has no understanding apart from what it is programmed to report.** Oxford University professor Nick Bostrom vividly illustrated this with his famous "paperclip maximizer" thought experiment. Imagine AI developers build an algorithm with the innocuous goal of collecting the maximum number of paperclips. The self-learning algorithm continuously finds new ways of collecting the clips. At first, the algorithm collects cartons of clips from the office supply closet; and then gathers misplaced clips lying under the sofa or desk in offices. To maximize the number of clips, it starts to manufacture clips from electrical duct metal and other galvanized steel, and it eventually melts all the metal on earth to manufacture paperclips.[4]

The experiment demonstrated how AI is a powerful tool but lacks the ability to understand context or values. AI does exactly what it is programmed to do—no more, no less. Even when AI "learns," it is doing so within the bounds of its programming. The hypothetical paperclip maximizer had been programmed to value paper clips and, without being told, it had no way of knowing that destroying humanity to make more paperclips was not desirable.

This feature of algorithms is **especially problematic for issues of diversity, equity, and inclusion** because certain groups may be less likely to be represented among developers who are giving AI "instructions." **The result may be potentially biased algorithms:** Some examples include a faulty facial recognition system that led to the arrest of an innocent person;[5] algorithmic models that may be more likely to recommend longer prison terms or potentially reject bail pleas for racially and/or ethnically diverse individuals;[6] or credit risk models that may be potentially biased against lending to people from certain ethnic and/or racial groups.[7]







## Process is the path forward

In order to achieve success, **finding a mechanism for cooperation that may allow each to share their strengths ethically, safely, and effectively is critical.** Having human workers act on unquestioned outputs of AI models is probably a bad process for cooperation. Get the cooperation process wrong and the result may quickly turn into a tragedy: a potentially biased AI algorithm making life-changing decisions about real people.

This is precisely the insight that chess grand master Garry Kasparov had when analyzing human-machine teams playing chess. He found that a "weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process."[8]

To support the improvement of the quality and equity of government decision making, a process should be created that gives human decision makers the benefit of AI's precision and consistency while confirming that AI is bounded by context-sensitive human values at every stage of its creation and operation.





# Using AI to debias noisy human decisions and deliver more equitable services

Biased AI algorithms can often be found in the headlines, but what is often missed is that purely human judgments in government are not without bias either. Years of research shows that human judgment is often affected by both "bias," and "noise," where bias is a systematic unfair outcome while noise is unwanted variability in judgments.[9]

## AI helping to improve human judgment

Variability in decisions is not inherently bad. In fact, there should be variability in decisions about two different loan applications from two different people, for example. **However, when that variability is not desired, it may create a problem.** For example, one study of juvenile courts found that judges were harsher in sentencing during the week following a loss by the local football team, and Black defendants were disproportionally affected by such judgments.[10] Nor are these variations in judgment limited to judges. A similar study indicated that immigration officers were less likely to grant asylum on hot days than on cooler days. Human decisions may vary depending on the time of the day, mood or weather.[11] Decisions that significantly affect human lives should not depend on who the judge is, whether it is hot or cold outside, or if a local football team lost a game.

**Government workers make scores of decisions where such variability is unwanted.** Childcare custody, bail decisions, and patent approvals may depend on which caseworker, judge, or examiner is assigned to make the decision. In such situations, the context-sensitivity of human judgment should be maintained, but insulated against too much variability from unrelated factors such as the weather or the quality of fare in the cafeteria. The consistency and precision of AI makes it an effective tool for this job. **While AI is consistent and precise, humans may be good at understanding varying contexts.** So together they may make an ideal team.[12]

## One additional ingredient to improving these sensitive decisions is setting a process by which humans and AI work together.

**One process that helps retain the context-sensitivity of human judgment and the consistency of AI is anchoring.** AI may provide an anchor that a human decision maker can adjust from. That **confirms AI isn't being too prescriptive and missing important factors unique to a case, and a human isn't being overly variable** based on a person's race or if they are hungry or not.

The potential of this human-AI teaming to improve government decision making may be seen in bail decisions. A team led by Sendhil Mullainathan, a professor at the University of Chicago's Booth School of Business, developed an AI model based on over 750,000 historical bail decisions to produce a probability of flight risk and broader effects on overall crime rates. For each case, the team had the same information as a judge: current crime, historical crimes, and accused's failure to appear before the court. The team also had information on whether the defendant was released or not, and after the release whether the individual failed to appear in court or was rearrested. The model's results indicated that crime could be reduced by 24.7% by focusing on high-risk individuals for pre-trial detention. The overall rate of detention for cases was unchanged. Alternatively, the number of people denied bail could be reduced by up to 42% by providing bail to people who are unlikely to recidivate with no overall increase in crime.[13]



# AI detecting bias and creating more equitable services

AI may not only reduce noise in human decisions. It also has the ability to also detect subtle inequities in service delivery. AI can then help shape solutions to deliver more equitable services.

### Detecting bias

Researchers from the University of Chicago used publicly available data from eight cities to predict future crimes one week in advance with nearly 90% accuracy. In another model, the team analyzed police response by examining arrests following a crime. The team found arrest rates increased in wealthier neighborhoods but remained flat or even dropped in poorer neighborhoods. In other words, **researchers had found a hidden pattern where more resources were allocated to wealthier areas** following crimes at the expense of poor neighborhoods.[14]

### Detecting bias and offering solution:

A researcher at the Illinois Institute of Technology and his former student analyzed Chicago's public transportation system to understand access to health care centers. The machine learning model developed by researchers identified inequities in far north and far south areas of Chicago, which had less than desirable access to health care centers.[15] The south side of Chicago in particular is home to many low-income residents. To mitigate inequity, the researchers developed a route optimization model for public buses so low-income neighborhoods could gain greater access to health care facilities. **The model increased access to medical facilities by 45% for people with minimal access** to such medical facilities without adding new buses.[16]




How teaming AI with humans can help debias decision-making

# Confirming the AI teammate is reliable

If agencies are using AI to improve human judgment, they should confirm AI is robust and reliable enough to not introduce additional errors of its own. To avoid such a scenario, the inputs of AI models—talent, data, and governance—need to be diverse.

## Diversity of data

The public sector increasingly uses AI in decision-making—from allocating resources to determining the eligibility of human services programs. However, **AI models are only as good as the datasets used to train them.**

**Organizations need to consider the suitability of variables included in their model.** For example, many lending institutions have traditionally used income and home ownership as variables to determine credit risks. But these variables may be systemically unfavorable to certain communities with historically low home ownership rates. **To avoid discrimination, lenders have included alternative data in conventional lending models.** They have also seen success with alternative variables such as deposit transactions, on-time payment of rentals and utilities to assess an applicant's ability to repay loan.[17]

**AI may help identify what alternative variables may be more effective.** Machine learning algorithms may potentially identify linkages not discovered through traditional credit scoring methods. In 2017, the Consumer Financial Protection Bureau allowed Upstart Network, Inc. to model underwriting and pricing decisions using both conventional and alternative data such as borrowers' education and employment history.[18] Upstart's approach saw loan approval rates increase by nearly 30% for some customer segments and credit costs decline by 15% to 17%. Such lending practices promote fair, non-discriminatory, and equitable access to credit, especially for groups of people whose poor credit history hinders their ability to access credit at reasonable rates.[19] What's more, these gains are realized at limited or even reduced risks to lenders.

## Diversity of talent

Historically, **many teams developing AI systems haven't been substantially reflective of broader society demographics.** For instance, as of 2021 only 22% of the global AI workforce is female.[20] Since bias sometimes originates from a narrow understanding of context where models operate, and AI only knows the context that its developers program, **building diverse teams that may see a problem from multiple perspectives is an important tool to mitigate potential bias.** To bring in diverse views, many organizations are including ethicists, sociologists, psychologists, and design thinkers to diversify AI teams and make AI more inclusive.

Take Bob.ai for example. In 2018, the Dallas Housing Authority partnered with the developer of Bob.ai app to help low-income voucher holders find rental units and assist authorities in inspections. Bejoy Narayana, an Indian immigrant and CEO of Bob.ai, was conscious that AI could further perpetuate segregation and redlining for minority communities, who were disproportionately relying on housing vouchers to pay their rents. So he, along with his racially and ethnically diverse team of developers, created an app so tenants could use the app without providing any identifying information.

By understanding the problem of housing for minority residents from different perspectives, **the Bob.ai team was able to identify novel solutions.** For example, one significant problem in the housing voucher program was a lack of available supply because many landlords did not want to participate in the program due to long wait times. The Bob.ai app automated workflows and communications for inspections, reducing the average wait time from 15 days to a single day to complete an inspection. This change attracted landlords who had not previously participated in the housing voucher program due to longer wait times, adding more than 20,000 units in the rental market. Greater supply has increased the chances of getting housing for low-income families.[21]





## Diversity in governance

Diversity of data and talent may reduce discrimination and de-bias algorithms, but **governance is also important to confirm that those controls work as designed and implemented.** All stakeholders, including the developers, end-users, and government executives should understand the importance of bias minimization in order to deliver inclusive AI services. But **driving fairness into the business processes of an organization takes structure,** so many organizations are creating new positions to oversee the governance of AI. The State of California is turning to its Chief Data Officer (CDO) to help it think through AI risks and data governance issues. Similarly, the Department of Defense brought in an AI ethics expert to leads the Responsible AI division of the Chief Digital and AI Office (CDAO).[22]

The exact structure of governance will likely vary by government agency. But their primary tasks should include putting in mechanisms to identify and mitigate potential bias and leading the response when potential bias is identified.

The governance group should also monitor algorithm audits and collaborate with external organizations such as regulators, policymakers, and algorithm accreditors to confirm responsible development of AI.[23]

# Bringing AI and humans together: Considerations for getting started

AI has the potential to reduce discrimination and provide inclusive government services, but only if human decision makers and AI work together to make the most effective, equitable decision possible (Figure 1).

## Using AI to improve the equity of government services

AI may reduce variability of subjective human judgments and mitigate potential human biases often shaped by individual and societal assumptions. Agencies may use AI to detect potential bias in individual judgments or even identify potential inequities in entire systems of government service. Several uses include:

**1 | Use AI to bring more and better information.** Organizations should take advantage of the fact that AI can ingest a significantly greater amount of data than a human could. As a result, human decision-makers may augment their own decisions with AI tools that identify patterns and make recommendations based on more effective data than a solely human judgment. For example, the United States Patent and Trademark Office has rolled out an AI component for a new patent search to assist its examiners in identifying the most relevant data to make decision as they examine over 600,000 application a year, averaging 10,000 words each.[24]

**2 | Adjust decision-making calculus.** AI teams should be trained to identify and reduce potential bias in existing public processes. For example, a cross-functional team of Deloitte professionals tested a public dataset on mortgage lending and found indications of disparate impact on minority races. To mitigate the potential bias, the team applied preprocessing bias mitigation techniques, such as variable repair (i.e., modification of variable distributions in the training dataset), to reflect fairer lending decisions without compromising the model's accuracy.[25]

**3 | Evaluate decision outcomes for equity.** Wherever data is sufficient, AI may be used to ascertain if human decisions are equitable. Often detection of potential bias or inequity is the first step toward addressing it, as was identified in the case of Chicago's health care facilities. Once researchers identified the access challenges to medical facilities for low-income residents, they developed an optimization model for buses that increased access to medical facilities.



How teaming AI with humans can help debias decision-making

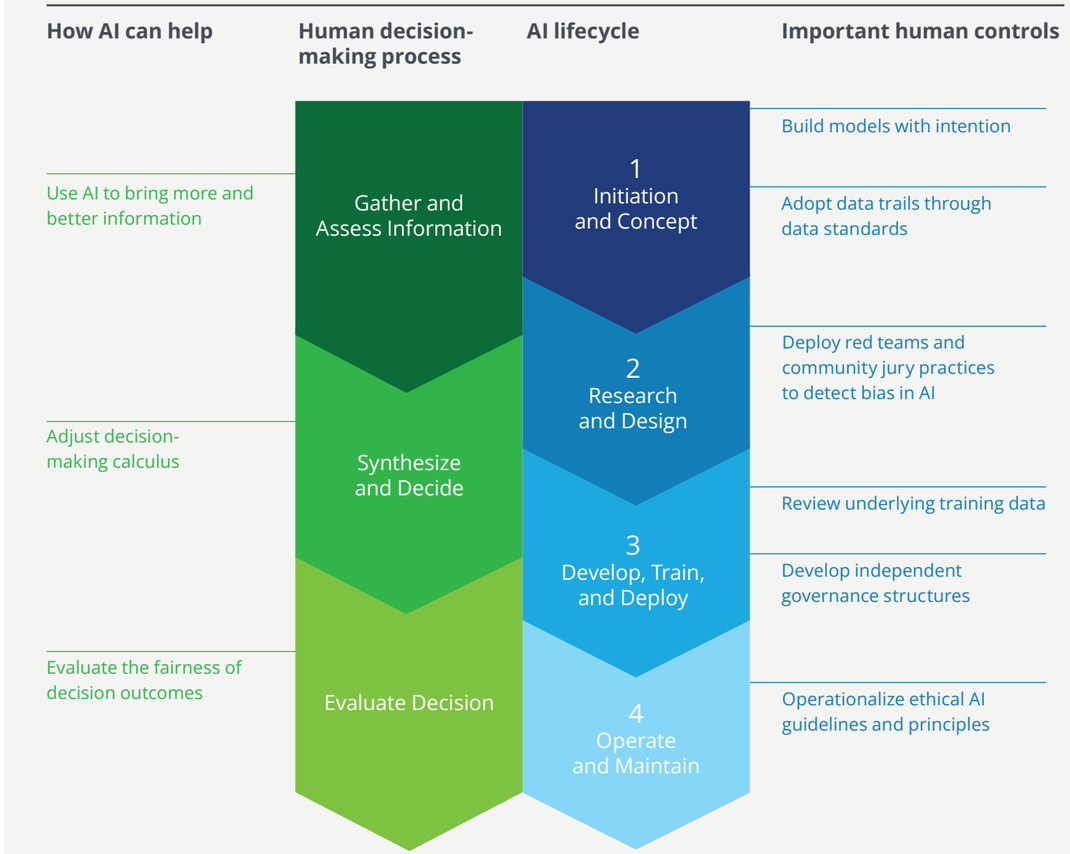



Figure 1.
Controls at each step of decision making can help bring the accuracy and breadth of AI to the context and judgment of human decision-making

# Confirming the fairness of AI

Agencies can take six steps to help confirm responsible development of AI and limit the implementation of algorithms that may result in individual and societal biases.

### Review underlying training data

One of the potential causes of algorithm bias is biased training data. **An algorithm is not biased in itself; however, the underlying training data may embed bias in an algorithm.** Agencies should ascertain whether data used to build the model accurately represents the relevant population. Additionally, even representative data may not be suitable for a given AI application. Government agencies should analyze datasets to identify potential historical biases before they are used as input for algorithms.[26] Using historical data sometimes requires extra care so as not to inadvertently propagate biases within that data.

### Adopt data trails through data standards

**Data standards, such as data cards, may give context** to the underlying data about how it was sourced or modified and its intended use, so developers may decide if a dataset is effective for their needs. Additionally, agencies should not only assess the accuracy of data labels but also ascertain if the data labeling exercise embedded any human biases in a training dataset.[27]

### Build models with intention

AI should be developed with clear objectives, intensive research of the processes being modeled, and outreach and stakeholder engagement. This approach should be motivated by practical cost-benefit considerations but also an organization's values and mission. **What AI should do or the metrics against which AI performance should be assessed are value judgments, not computational exercises.** In practice, this means that rather than data scientists simply working from available data, AI talent and organizational leadership should take the time to understand and map how they believe relevant inputs and outputs are linked and how they influence AI design. A near-term step teams may take is to select the outcome of interest being modeled and interrogate how it relates to the underlying process and overarching mission values.[28]



How teaming AI with humans can help debias decision-making

### Deploy "red teams" and community jury practices to detect potential bias in AI

**Red teaming is a practice where internal and external teams use AI to cause harm with the intention of exposing risks of models.** The red team may also detect behavioral errors like group-think or confirmation bias. Community juries, borrowed from the idea of citizen juries, also allow developers to test adverse impacts of AI. **The aim is to bring representative people from diverse populations, especially from historically underrepresented communities, to get their perspectives on how a model may adversely affect them.**[29] To tap into the wisdom of crowd, the Department of Defense and the Department of Homeland Security have launched 'bug bounties' to identify security vulnerabilities in their systems; agencies may consider launching 'bias bounties' on the similar lines to detect potential bias in AI systems.[30]

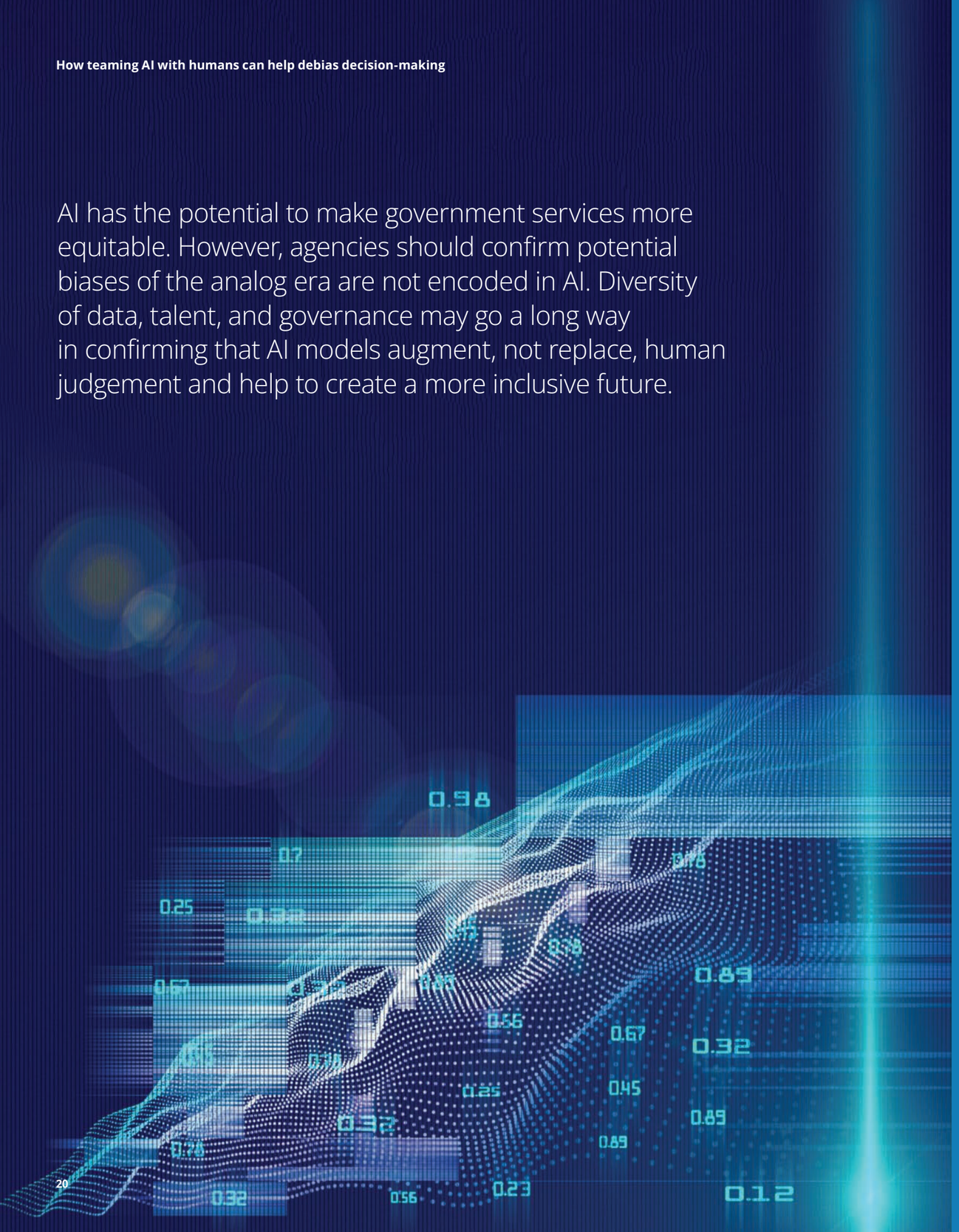
### Develop independent governance structure

Like the risk management or internal audit function of financial institutions, agencies should keep the AI governance structure at arm's length from business functions. **Independent structures allow agencies to deconflict with business decisions and build effective control mechanisms** to prevent negative impacts of AI.[31]

### Operationalize ethical AI guidelines and principles

**Agencies should move beyond just releasing ethical guidelines and frameworks and put them into practice.** Establish a governance structure to govern the development of ethical AI, cultivate an organizational culture that nurtures ethical AI, and continuously monitor AI to confirm models are not generating potentially biased outputs. Like financial statements, third-party audits of AI models may help confirm bias does not go undetected for long.[32]

AI has the potential to make government services more equitable. However, agencies should confirm potential biases of the analog era are not encoded in AI. Diversity of data, talent, and governance may go a long way in confirming that AI models augment, not replace, human judgement and help to create a more inclusive future.






How teaming AI with humans can help debias decision-making

# Endnotes


1 IALEIA: Psychology of Intelligence Analysis

2 ibid.

3 Deloitte Insights: Realizing the full potential of AI in the workplace

4 Quartz: The humble office-supply item that can explain humanity's imminent doom

5 The New York Times: Wrongfully Accused by an Algorithm

6 Brookings: Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms

7 Brookings: Reducing bias in AI-based financial services

8 New York Review: The Chess Master and the Computer

9 Daniel Kahneman, Olivier Sibony and Cass R. Sunstein, *Noise: A Flaw in Human Judgment*. New York, Little, Brown Spark, 2021.

10 American Economic Journal, Applied Economics: Emotional Judges and Unlucky Juveniles

11 Harvard Business Review: Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making

12 Deloitte Insights: Superminds, not substitutes

13 Stanford Computer Science: Human Decisions and Machine Predictions

14 University of Chicago, Biological Sciences Division: Algorithm predicts crime a week in advance, but reveals bias in police response

15 Illinois Tech: Model Transportation: Using Math to Maximize Public Bus Routes and Increase Access to Health Care Centers

16 Association for Computing Machinery: A public transit network optimization model for equitable access to social services

17 Consumer Finanical Protection Bureau: An update on credit access and the Bureau's first No-Action Letter

18 ibid.

19 ibid.

20 The Print: Only 22% women in AI jobs—The gender gap in science and technology, in numbers

21 The Pew Charitable Trusts: Programmers and Lawmakers Want AI to Eliminate Bias, Not Promote It

22 Fed Scoop: Pentagon names new chief of responsible artificial intelligence

23 Chicago Booth: Algorithmic Bias Playbook

24 CIO: USPTO takes human-first approach to AI innovation

25 Deloitte Insights: Trustworth open data for trustworthy AI

26 Deloitte Insights: AI model bias can damage trust more than you may know. But it doesn't have to

27 Deloitte Insights: Trustworth open data for trustworthy AI

28 The New York Times: Biased Algorithms Are Easier to Fix Than Biased People

29 InfoWorld: What is AI bias mitigation, and how can it improve AI fairness?

30 Security Week: DoD Launches 'Hack US' Bounties for Major Flaws in Publicly Exposed Assets; Department of Homeland Security: DHS Announces "Hack DHS" Bug Bounty Program to Identify Potential Cybersecurity Vulnerabilities

31 Deloitte: Developing and deploying trustworthy AI in Government

32 Deloitte Insights: Trustworth open data for trustworthy AI

# Reach out for a conversation.


**Tasha Austin**
Advisory Principal
Deloitte & Touche LLP
Director
Deloitte AI Institute
for Government
**laustin@deloitte.com**

**Joe Mariani**
Senior Research Manager
Deloitte Services LP
**jmariani@deloitte.com**

**Devon Dickau**
DEI Consulting Services Leader
Deloitte Consulting LLP
**ddickau@deloitte.com**

**Pankaj Kamleshkumar Kishnani**
Researcher
Deloitte Services LP
**pkamleshkumarkish@deloitte.com**

**Thirumalai Kannan**
Researcher
Deloitte Services LP
**tkannand@deloitte.com**




**Deloitte.**