Bias in the Loop

How Human Judgment Shapes AI-assisted Anti-fraud Work

By Gavin Ugale

In 1983, the British engineer Lisanne Bainbridge wrote in *Ironies of Automation* that "the more advanced a control system is … the more crucial may be the contribution of the human operator."[1] Her argument was not just that humans have a role to play relative to automated systems; it was that automating a process changes the cognitive demands on the humans in charge of those systems. Over time, hands-on operating becomes monitoring, and skills atrophy from lack of use. And when systems fail, human operators must intervene with eroded skills at a moment when cognitive demands are highest. Four decades later, this dynamic illustrates the behavioral risks agencies face as they design and deploy artificial intelligence (AI) systems, including those aimed at strengthening fraud prevention and detection.[2] AI is not the industrial control system Bainbridge studied, but the cognitive dimension of her argument carries over to the challenges agencies face today.

Various federal policies and standards governing use of AI in government call on agencies to keep a human in the loop (HITL) to supervise AI systems and ensure their accountability. That supervision is critical, but the traditional framing of HITL makes it easy to overlook the human and machine interactions that influence how well AI systems perform, including the cognitive dimension Bainbridge identified. AI in the loop (AITL) is a complementary idea that shifts the focus from humans as overseers of AI to humans as users, with AI playing a supporting role in a human-led process, as shown in **Figure 1**. It helps to better illustrate how fraud examiners and auditors think and decide when using AI, and how their biases can interact with AI outputs.

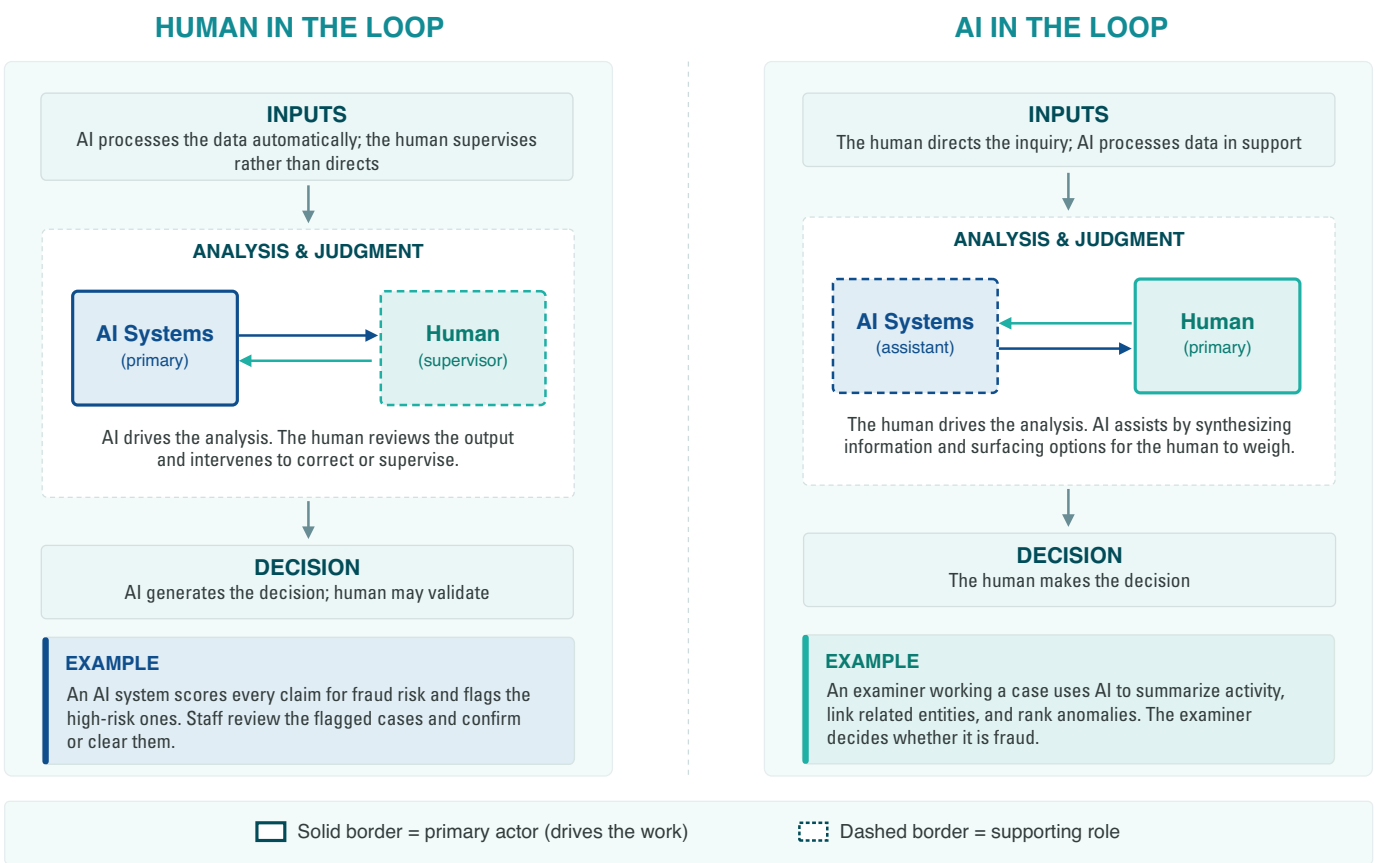
Agencies that disregard this interaction as part of their approach to AI governance risk deploying systems that amplify these biases in ways that could otherwise be mitigated. From this vantage, this article explores cognitive biases that can manifest in AI-assisted anti-fraud work and their interplay with institutional factors during deployment. With this understanding, we can consider several practical recommendations for how federal agencies can account for human cognition when adopting AI.

COGNITIVE BIASES IN ANTI-FRAUD ACTIVITIES AND AI

Studies and post-mortems on AI failures often focus on upstream bias, particularly algorithmic bias that becomes embedded during design and development of AI systems. The National Institute of Standards and Technology (NIST) refers to this as statistical and computational bias: errors that arise when training data is not representative of the population, occurring "in the absence of prejudice, partiality or discriminatory intent."[4] The Dutch childcare benefits scandal illustrates the extreme consequences of this kind of bias coupled with institutional failures. A discriminatory risk model employed by the Dutch Tax and Customs Administration between 2005 and 2019 contributed to tens of thousands of families being falsely accused of benefits fraud, ultimately leading to the resignation of the government.[5]

Behavioral mechanisms and cognitive biases during deployment receive less attention in the AI and anti-fraud literature than upstream algorithmic bias, but they are just as consequential. Unlike statistical and computational bias, which can be inherited from a vendor's model, these biases arise within agencies as frontline staff, risk managers and auditors work with AI systems to carry out different anti-fraud tasks. An AI-generated fraud alert, risk score or document summary requires a human reviewer to interpret an output and assess its reliability. How they do this varies by output type. An analyst checking an AI-generated risk score against the features that produced it draws on different skills than one reviewing a generative AI summary against lengthy source materials.

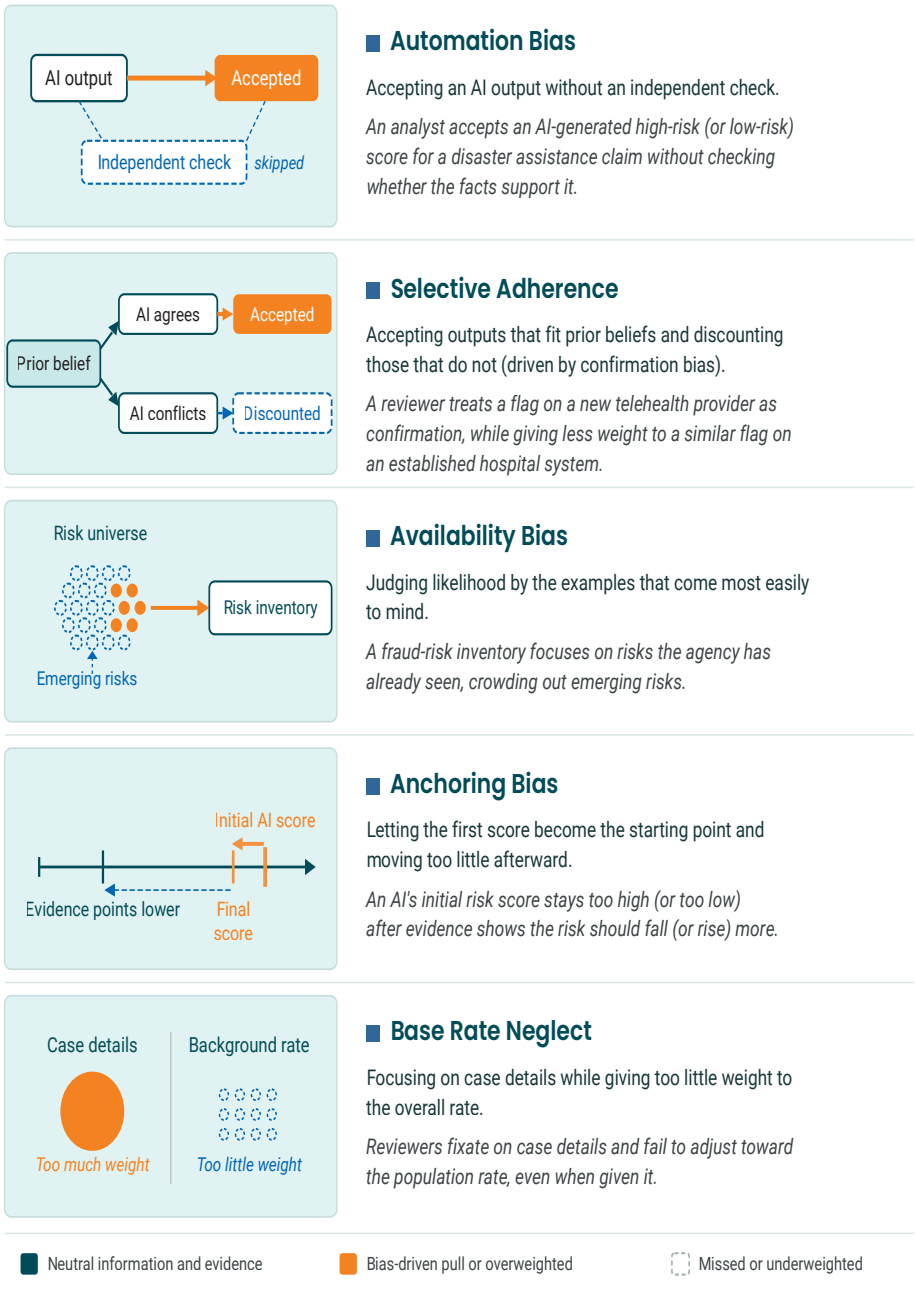
Figure 1. Comparison Between HITL and AITL



Source: Adapted from Natarajan et al., 2025[3]

There are few agency reports or audits that directly examine how human judgment and cognition interact with AI, let alone in the anti-fraud context. These issues, including considerations about human cognition and bias, mostly appear in standards and manuals, such as guidance for auditors on preserving professional skepticism.[6] Academic research at the intersection of AI and fraud prevention in government is equally sparse, but fields like behavioral economics, cognitive psychology, auditing and human-computer interaction offer insights into biases that this article highlights and are relevant to anti-fraud work. While not exhaustive, **Figure 2** presents five prominent biases and how they play out in AI-assisted anti-fraud work, each covered in greater detail below.

Figure 2. Five Deployment-phase Cognitive Biases With Examples From AI-assisted Anti-fraud Work



Source: Author

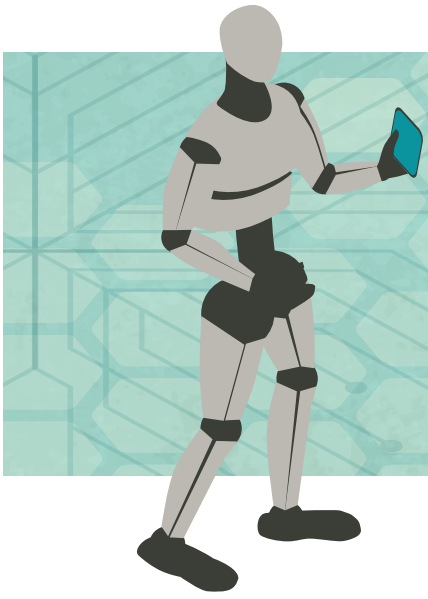
Automation Bias

Consider an analyst reviewing disaster assistance claims. An AI system flags a claim as high-risk for fraud, and the analyst accepts the score and refers the case for investigation without checking whether the flag aligns with the claimant's actual circumstances. This is *automation bias* at work, which is the tendency for people to over-rely on outputs from automated systems (a closely related phenomenon, automation complacency, refers to under-monitoring of those systems).[7] Human factors research shows that teaching people about biases in advance helps less than interventions applied at key decision points, and people often defer to AI-generated recommendations even when they are wrong.[8]

Attaching explanations to the output, such as showing why the system produced a high score, does not necessarily help to mitigate the bias. In fact, it can increase deference to an AI output when it is incorrect, because of people's tendency to form an overall judgment about a system's reliability rather than evaluate each result or recommendation on its own merit.[9] Some research suggests over-reliance can also reflect a strategic response to the costs and benefits an examiner faces in a given task.[10] For instance, if checking a flag drains resources, such as the time and effort required for cross-checking sources, accepting an AI output can cost less than verifying it. By comparison, *algorithmic aversion*, a preference for human advice over an algorithm's, can lead to under-reliance on AI, as shown in one study of auditors who discounted the same evidence when it came from AI rather than a human specialist.[11]

Selective Adherence

Now consider an auditor who believes that new telehealth providers represent a concentrated fraud risk. An AI system flags billing anomalies across a sample of providers, and the auditor pursues a flag on a new applicant as confirmation of that belief, while a comparable flag on



Agencies' success in deploying AI systems for fraud prevention and detection depends on recognizing that cognitive biases will always be in the loop, regardless of whether they subscribe to an AITL or HITL framing of the role of humans.

an established hospital system gets less scrutiny. This is a behavioral pattern called *selective adherence*, when decision-makers accept AI-generated outputs that align with their pre-existing beliefs and discount those that do not.[12] *Confirmation bias* is the underlying mechanism, well-documented in the audit literature and addressed in government auditing standards on professional judgment.[13]

Research shows selective adherence occurs whether the advice comes from AI or from humans, with no significant difference between the two, so replacing human advice with algorithmic advice does not eliminate the bias.[14] The Dutch childcare benefits scandal shows how selective adherence compounds algorithmic bias at scale. The tax authority's algorithmic risk model embedded discriminatory features, such as nationality, while frontline human reviewers reinforced that bias rather than correcting it. This is also why the AITL framing is useful. Keeping the analytical focus on the human user is what allows agencies to identify biases and find ways to counter them.

Availability Bias

Fraud risk assessments offer a structured process for further exploring the potential influence of bias on human judgment and decision making. Imagine a team developing a fraud risk inventory as part of the assessment process. They might convene a workshop, organize focus groups, pull reports and historical case data, or draw on institutional memory to catalog the risks the program faces. Say they succeed in building a comprehensive risk inventory, but it primarily reflects the schemes they know well, leaving out emerging ones. This is *availability bias*, the tendency to judge how likely something is by how easily examples come to mind.[15] Here it narrows the universe of risks to what the agency has already observed.

Recency bias, when newer fraud schemes are easier to recall, works the same way by shaping the inventory around recent memory rather than the full risk landscape. Unaware of these blind spots, an agency can embed bias at this stage through design choices like biased alert thresholds, which then propagate downstream.

Anchoring Bias

In AI-assisted fraud risk assessments, the AI output can introduce a bias of its own. The classic approach involves scoring inherent risks (before controls) and residual risks (after controls) to gauge how well controls reduce the agency's exposure to fraud risks. Scoring risks varies in complexity, but it typically relies on analyzing features like likelihood, impact and velocity. Now suppose an AI system generates the inherent score, pulling from historical data and program characteristics. That number flashes on a dashboard for the examiner before they assess the effectiveness of controls, and then the residual score lands only slightly below that, even when an independent review would justify a much lower residual risk score. This is an example of *anchoring bias*, which is the tendency to rely too heavily on an initial value or estimate and to adjust insufficiently when new information warrants change.[16] The AI's inherent risk score becomes the anchor, making it difficult for the examiner to credit the controls as much as the evidence warrants.

Federal agencies are already rolling out AI systems to support fraud prevention and detection activities in ways that can anchor judgment. The Internal Revenue Service (IRS) uses AI to score and classify tax returns as low-risk or high-risk for potential fraud, as the U.S. Government Accountability Office (GAO) highlighted in a 2026 report on the IRS's use of AI.[17] Sorting returns at this stage is a sound triage strategy. The potential for bias to enter this process comes at the next stage, when IRS examiner reviews the AI's output for a flagged return and decides whether to refer it for further investigation. That process involves human judgment, which is the point when an anchor, like a high-risk designation or score, can take hold and pull the decision toward the conclusion the AI already reached. Research indicates this happens even when subsequent information warrants a different conclusion.

Base Rate Neglect

While anchoring is overreliance on an initial value, *base rate neglect* operates in the other direction when people underuse a reference point, like how often fraud occurs across a population.[18] In a classic experiment,

auditors estimating the likelihood of fraud fixated on details of the company in front of them. Even when researchers gave them the actual rate of fraud across the broader population, they failed to sufficiently adjust their estimates toward it.[19] This error in judgment can push an estimate too high when fraud is rare, or too low when fraud is more common. Base rate neglect is one way evidence-weighting goes wrong. *Averaging bias* is another, which can lead to diluting strong risk indicators when assessed alongside weaker evidence at the same time.[20]

CONDITIONS THAT SHAPE HUMAN-AI INTERACTION

Biases do not occur in isolation. How an agency presents its AI outputs, and the conditions its examiners work under, shape how strongly each one takes hold. Show the AI system's recommendation with an explanation attached, and acceptance rises regardless of whether the AI system is correct.[21] Pin a risk score to the top of a dashboard and it acts as an anchor for whatever the examiner reads next. In addition to these choices, environmental factors can shape the interpretation of AI outputs. When people are pressed for time, they are more likely to accept suggestions of the AI system.[22] Time pressure becomes especially acute during disaster or emergency spending, when agencies prioritize speed of payments over anti-fraud safeguards. Automation bias, the risk of deferring blindly to an AI output, can be amplified under these conditions.

Coordination is another condition that shapes how well AI tools support fraud work. When it breaks down or is absent, it leaves more room for bias to influence outcomes. Standards for managing fraud risks call for a feedback loop so that the people designing preventive controls can learn from the outcomes of investigations and adjudicated cases. In practice, agencies struggle to operationalize this. GAO recently found that the U.S. Department of Defense (DOD) does not obtain or assess information on adjudicated procurement fraud cases from its own Office of the Inspector General, among other sources. As a result, DOD's fraud risk assessments lack critical information about the salience and potential impact of certain types of risks.[23]

Poor coordination and missing feedback loops can directly raise the risk of bias. When examiners lack new information about the risk universe, they are more likely to fall back on their heuristics and prior beliefs. The same missing feedback also makes it more difficult to assess the quality of the AI system they are using. Research shows that, without this signal, people gauge a system's reliability by how often it agrees with their own judgment. That invites two errors: overreliance on the AI system when it mirrors their biases, and under-reliance on it when it would correct them.[24]

MITIGATING BIASES IN AI DEPLOYMENT

More attention to cognitive biases when designing, deploying, and auditing AI systems to prevent and detect fraud is critical for ensuring that taxpayer money invested in these AI systems is well spent. It is also essential for avoiding discriminatory and damaging societal impacts, such as those that materialized in the Dutch childcare benefits scandal. Yet how humans interact with AI systems is often overlooked, because the focus in policy and practice is primarily on how humans oversee these systems as an internal control. That very process of overseeing is itself vulnerable to biases.

Agencies' success in deploying AI systems for fraud prevention and detection depends on recognizing that cognitive biases will always be in the loop, regardless of whether they subscribe to an AITL or HITL framing of the role of humans. Fortunately, unlike the algorithmic bias embedded in a system upstream, these deployment-phase biases sit within the sphere of control and influence of different agencies. The following actions offer a starting point to identify biases and deliberately mitigate them from the perspective of deploying agencies, oversight bodies and policy setters, like the Office of Management and Budget. While the focus is on anti-fraud work, several ideas apply wherever agencies use AI to inform decisions.

When Piloting AI Systems, Agencies Should Account for Behavioral Patterns and Biases, Not Just Technical Accuracy

A useful pilot measures more than whether a new AI system is accurate. It measures how users interact with the system and the quality of the decisions they make with it. A behaviorally-informed pilot tracks whether users defer to the tool when it is wrong, whether they catch it when it makes mistakes and whether time pressure or workload changes either user action. A pilot that ignores these considerations will provide an incomplete picture of whether the tool will work at scale in the hands of busy examiners whose judgment is susceptible to bias. These findings can also serve as a starting point for documenting an AI system's behavioral risks and consistent monitoring of these risks over time.

Findings from the pilot can also help shape the design of an AI system and its output to counter bias. One countermeasure for bias is called a *cognitive forcing function*, a small design choice that makes the reviewer engage with the output before accepting it. Evidence suggest that asking people to commit to their own judgment before seeing the AI's, and showing the latter only upon request, can reduce over-reliance.[25] In practice, this could mean having a fraud examiner record their own conclusion before an AI score appears, so the number cannot serve as an anchor, or delivering the score upon request rather than pinned to the top of the dashboard. Agencies could also require a brief reason to accept or override a flag, which targets selective adherence. Research indicates that reviewers may resist added effort, so agencies should apply behavioral controls judiciously to avoid overburdening the process.

Oversight Bodies Can Expand Their Focus to Include the Behavioral Lens

Oversight bodies can apply the same behavioral lens as agencies when auditing and evaluating AI systems. Auditors can examine a model's accuracy, its governance, and the presence of human supervision, yet still miss behavioral patterns

(e.g., anchoring on scores or deferring to AI under time pressure) that undermine the AI system's objectives and lead to inefficient or wasteful use of taxpayer money. Existing standards and research provide a foundation for this approach. GAO's AI Accountability Framework already addresses bias, although largely with a focus on data as opposed to how people interact with an AI system and interpret its outputs.[26] Behavioral considerations and audit questions that concentrate on human-AI interactions would be a natural extension of it. A high acceptance rate of an AI output, for instance, is consistent with an accurate model as well as automation bias. An auditor could test for automation bias by examining whether reviewers who accept nearly every AI flag are independently checking those cases, and whether acceptance climbs as case volume rises.

Auditors themselves are not exempt from these biases. Generally Accepted Government Auditing Standards already require them to exercise professional skepticism, including a questioning mind and a critical assessment of evidence when conducting an engagement. AI-generated outputs will increasingly become that evidence. Calibrated reliance — trusting a model when it is right and overriding it when it is not[27] — is what professional skepticism requires when the evidence is machine-generated. Making this concept explicit, whether in trainings, guidance or future updates to auditing standards, would extend an already prescribed standard to a new kind of evidence.

Future Iterations of Federal AI Policy and Guidance Could Elaborate on the Behavioral Dimension

Under current federal policy, agencies must maintain public inventories of their AI use cases and update them annually, recording each system's purpose, outputs and underlying data. There is no federal mandate to document behavioral risks, and the recent trend has been to streamline rather than expand what agencies report, which makes a near-term requirement unlikely. Absent one, many agencies will not. Nonetheless, standard setters such as NIST explicitly identify human-AI interactions, including overreliance and automation bias, as a risk category.[28]

Agencies would not need to start from scratch. The process for building AI use case inventories is an established channel for articulating behavioral considerations, including the bias risks of specific systems and agencies' plans to mitigate them. Agencies' own pilots and testing, coupled with a growing body of academic literature, provides a foundation for understanding the biases and errors of judgment that can affect AI-assisted anti-fraud tools and other applications. For now, documenting how human judgment can affect, or be affected by, an AI system is the agency's choice. For stewards of taxpayer money, it is the responsible one.

CONCLUSION

Bainbridge identified an earlier version of this human-machine dynamic four decades ago, and it remains pertinent today. Automating a process did not make the human operators she studied obsolete. It gave them the harder job of dealing with system failures after the skills to do so eroded from lack of use. She observed that it is "impossible for even a highly motivated human being to maintain effective visual attention" on tasks where little happens.[29] AI-assisted anti-fraud work faces the same types of risks today. The judgment applied to a risk score or a flagged claim remains human, and so do the biases that shape it. Seeing AI as part of existing human-led processes is what allows agencies to identify those biases and mitigate them. Whether AI strengthens fraud prevention or quietly entrenches existing biases will depend not just on the design of the underlying model but on the people in the loop, and on how proactive agencies are in preparing them.

Endnotes

1. Bainbridge, Lisanne, "Ironies of Automation," *Automatica*, Volume 19, Issue 6, November 1983, pp. 775–779.
2. In this article, "AI systems" refers primarily to classification-style tools that produce discrete outputs (e.g., risk scores, anomaly flags, binary classifications). This reflects the focus of the empirical literature cited here and aligns with many of the public examples of AI-assisted federal fraud work. The themes discussed also apply to generative AI, but the evidence on these tools is still emerging, and insights from other types of AI may not carry over to it.
3. Natarajan, Sriraam, Saurabh Mathur, Sahil Sidheekh, Wolfgang Stammer and Kristian Kersting, "Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?," Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, 39(27).
4. Schwartz, R., A. Vassilev, K. Greene, L. Perine, A. Burt and P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," NIST Special Publication 1270, NIST, March 2022.
5. Heikkilä, Melissa, "Dutch scandal serves as a warning for Europe over risks of using algorithms," *Politico*, March 29, 2022.
6. Searches conducted with the Program Integrity Alliance's GovQuery search engine, which covers GAO and OIG reports published since 2010.
7. Goddard, K., A. Roudsari and J. C. Wyatt, "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," *Journal of the American Medical Informatics Association*, Volume 19, Issue 1, January 2012, pp. 121–127.
8. Buçinca, Z., M. B. Malaya and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proceedings of the ACM on Human-Computer Interaction*, Volume 5, Issue CSCW1, April 2021, pp. 1–21.
9. Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro and Daniel S. Weld, "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance," *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021, Article 81, pp. 1–16.
10. Vasconcelos, H., M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein and R. Krishna, "Explanations Can Reduce Overreliance on AI systems During Decision-making," *Proceedings of the ACM on Human-Computer Interaction*, Volume 7, Issue CSCW1, pp. 1–21.
11. Commerford, B. P., S.A. Dennis, J. R. Joe and J.W. Ulla, "Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence," *Journal of Accounting Research*, Volume 60, Number 1, March 2022, pp. 171–201.

12. Alon-Barkat, S. and M. Busuioc, "Human-AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice," *Journal of Public Administration Research and Theory*, Volume 33, Issue 1, January 2023, pp. 153–169.
13. GAO, "Government Auditing Standards 2024 Revision," GAO-24-106786, Feb. 1, 2024.
14. See Endnote 12.
15. Tversky, A. and D. Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, Volume 5, Issue 2, September 1973, pp. 207–232.
16. Kahneman, D, "Thinking, Fast and Slow," Farrar, Straus and Giroux, New York, October 2011.
17. GAO, "Artificial Intelligence: IRS Actions Needed to Address Skills Gaps, Information Quality, and Strategic Management," GAO-26-107522, March 24, 2026.
18. Joyce, E. J. and G.C. Biddle, "Are Auditors' Judgments Sufficiently Regressive?," *Journal of Accounting Research*, Volume 19, Number 2, Autumn 1981, pp. 323–349.
19. Ibid.
20. Lambert, T. A. and M. Peytcheva, "When Is the Averaging Effect Present in Auditor Judgments?," *Contemporary Accounting Research*, Volume 37, Issue 1, May 2019, pp. 277–296.
21. See Endnotes 8 and 9.
22. Swaroop, S., Z. Buçinca, K.Z. Gajos and F. Doshi-Velez, "Accuracy-Time Tradeoffs in AI-assisted Decision Making Under Time Pressure," *IUI '24: Proceedings of the 29th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, April 5, 2024.
23. GAO, "DOD Fraud Risk Management: DOD Should Expeditiously and Effectively Implement Fraud Risk Management Leading Practices," GAO-25-108500, June 2025.
24. Lu, Z. and M. Yin, "Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks," *CHI '21*, May 2021, Article 78, pp. 1–16.
25. See Endnote 8. The trade-off noted here, that the designs reducing overreliance the most were rated least favorably by users, is reported in the same study.
26. GAO, "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities," GAO-21-519SP, June 30, 2021.
27. Lee, J. D. and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors*, Volume 46, Issue 1, Spring 2024, pp. 50–80.
28. NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, July 2024.
29. See Endnote 1.

Gavin Ugale is the executive director and cofounder of the Program Integrity Alliance. He has more than 20 years of experience supporting governments, both in the U.S. and internationally, to strengthen the integrity and accountability of public spending. Previously, Gavin worked in Paris, France at the Organization for Economic Cooperation and Development (OECD) for nearly a decade, where he pioneered efforts to help anti-corruption and oversight bodies modernize through the strategic use of technology and data, and developed and monitored global standards such as the OECD's Recommendation on Public Integrity. Before the OECD, Gavin served at GAO where he was the primary author of its Framework for Managing Fraud Risks in Federal Programs, which was later incorporated into the Fraud Reduction and Data Analytics Act.