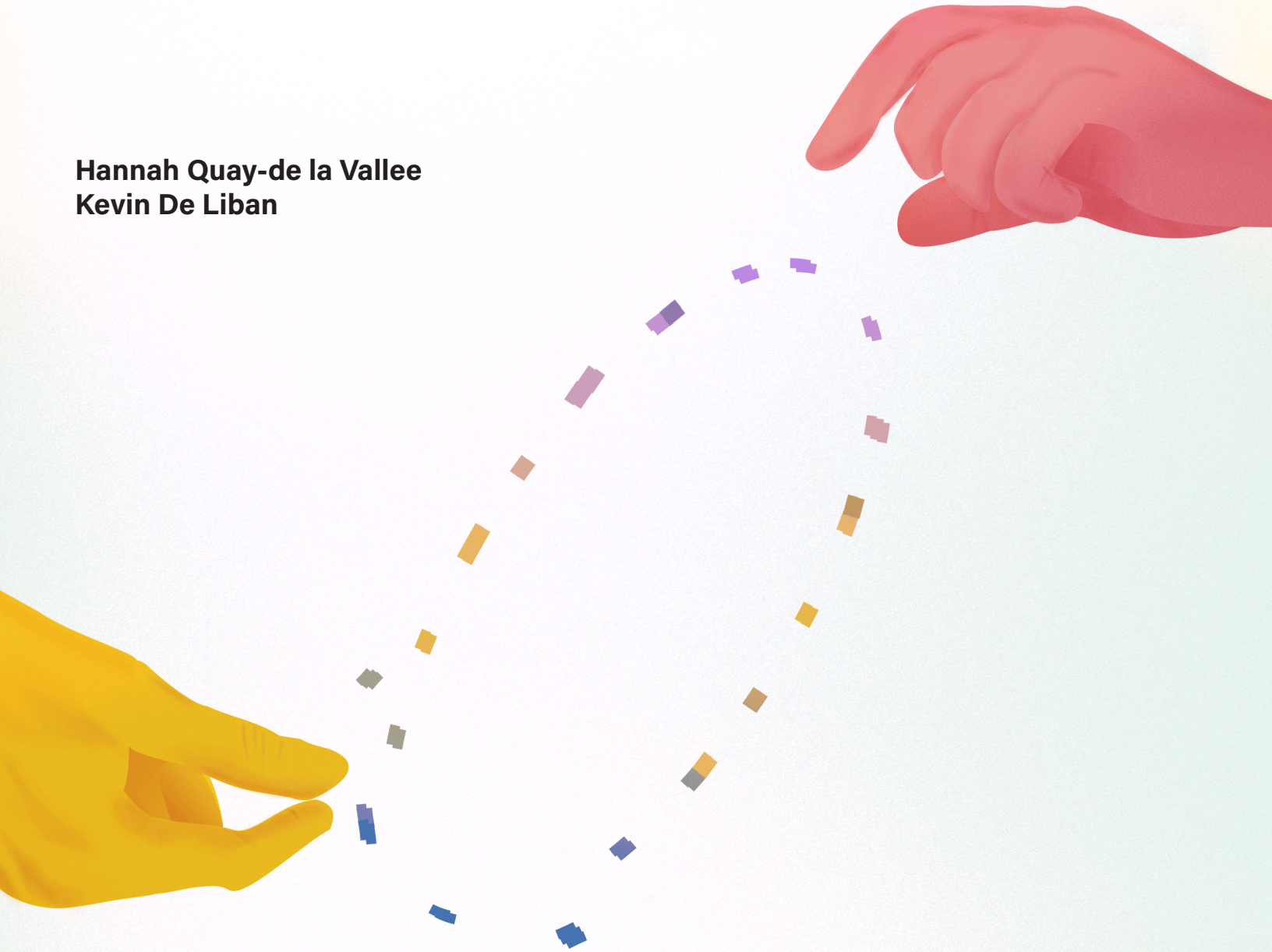
SEPTEMBER 2025

Rethinking the Loop

Encircling Public Benefits AI with Human Oversight

Hannah Quay-de la Vallee
Kevin De Liban



CENTER FOR DEMOCRACY & TECHNOLOGY

BENEFITS TECH ADVOCACY HUB

CENTER FOR DEMOCRACY & TECHNOLOGY

The **Center for Democracy & Technology (CDT)** is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

As governments expand their use of technology and data, it is critical that they do so in ways that affirm individual privacy, respect civil rights, foster inclusive participatory systems, promote transparent and accountable oversight, and advance just social structures within the broader community. **CDT's Equity in Civic Technology Project** furthers these goals by providing balanced advocacy that promotes the responsible use of data and technology while protecting the privacy and civil rights of individuals. We engage with these issues from both technical and policyminded perspectives, creating solutions-oriented policy resources and actionable technical guidance.

BENEFITS TECH ADVOCACY HUB

The **Benefits Tech Advocacy Hub** is a project of the National Health Law Program, TechTonic Justice, and Upturn to nurture a community of frontline advocates to fight technology-enabled cuts to public benefits and foster collective efforts to promote public benefits systems that meet people's basic needs.

SEPTEMBER 2025

Rethinking the Loop

Encircling Public Benefits AI with Human Oversight

Hannah Quay-de la Vallee
Kevin De Liban

This report was co-authored by the Benefits Tech Advocacy Hub and CDT.

This report was informed by a workshop designed to solicit the expertise of advocates, government officials, legal and technical experts, and recipients of public benefits. Their contributions were invaluable to the creation of the report.

With contributions from Tim Hoagland on layout and illustration.

References in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

BENEFITS TECH ADVOCACY HUB

CENTER FOR DEMOCRACY & TECHNOLOGY

Contents

01 Executive Summary

Artificial Intelligence (AI) and other algorithmic systems[1] **are becoming increasingly commonplace in the provision of public benefits, performing tasks ranging from administrative data entry to determinations about program eligibility and allocation of care.** However, not only do these systems often fall short of their goals, they frequently cause harm to the very people that public benefits systems exist to support. These harms include financial impacts like ruined credit, social impacts like decreased independence and loss of community, and medical impacts like worsened health and preventable death. [2]

The most effective way to limit the harm of algorithmic decision making is to consider whether an algorithm is appropriate

1 As explained on p. 13, This report defines "AI and other algorithmic systems" broadly to encompass technologies with varying levels of sophistication.

2 Rachael Kohl, *Automated stategraft: Faulty programming and improper collections in Michigan's unemployment insurance program*, Wisconsin Law Review (Apr 22, 2024) https://wlr.law.wisc.edu/automated-stategraft-faulty-programming-and-improper-collections-in-michigans-unemployment-insurance-program/ [https://perma.cc/ZU7A-539U];
Colin Lecher, *What happens when an algorithm cuts your health care*, The Verge (Mar 21, 2018) https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy [https://perma.cc/2UAZ-C4RL];
Children's Defense Fund Texas, *In harm's way: True stories of uninsured texas children* (Apr 2, 2007) https://www.childrensdefense.org/cdf-releases-new-report-in-harms-way-true-stories-of-uninsured-texas-children/ [https://perma.cc/8TBZ-ECBL; https://perma.cc/V7XK-QEES].

for the situation and, if not, avoid using it. Even with appropriate uses, human oversight is required. In theory, this means the organization or individual overseeing the system has actual power to intercede in the system's operations and decision making, including monitoring, modifying, suspending, or ending its use.[3] In some cases, there are legal mandates to provide this oversight,[4] but even in the absence of these requirements, human oversight is considered one of several requisite and standard practices to promote accountability.

Human oversight is considered one of several requisite and standard practices to promote accountability.

Although there is a dearth of research and guidance about what human oversight of AI should look like in practice, this brief draws upon existing scholarship, policy expertise, and experience challenging unjust uses of AI to lay out best practices for an effective and inclusive human oversight framework. These best practices span several different components of oversight:

- **Take an inclusive view when determining how and when people should be engaged.** Humans should be involved in *individual* decisions made by AI. The full range of people likely to be impacted by AI systems, particularly current recipients of benefits, need to have meaningful input into decisions *about* AI, including in its adoption, design, use, and oversight. This means understanding when impactful decisions are being

3 Gina M. Raimondo & Laurie E. Locascio, *AI Risk Management Framework*, National Institute of Standards and Technology (Jan 2023) https://airc.nist.gov/airmf-resources/airmf/ [https://perma.cc/Q96M-39RU; https://perma.cc/NK96-SRQE];
City of Seattle, *Responsible Artificial Intelligence (AI) Program – Generative Artificial Intelligence Policy* (accessed Jun 16, 2025) https://www.seattle.gov/tech/data-privacy/the-citys-responsible-use-of-artificial-intelligence [https://perma.cc/5S8C-4V28; https://perma.cc/AC7A-QWR7];
Jeff Maxon, *Generative Artificial Intelligence Policy*, Kansas Office of Information Technology Services (July 25, 2023) https://www.governor.ks.gov/home/showpublisheddocument/405/638744386434630000 [https://web.archive.org/web/20250307073659/https://www.governor.ks.gov/home/showpublisheddocument/405/638744386434630000].

4 Colorado General Assembly, *SB22-113 Artificial intelligence facial recognition* (2022) https://leg.colorado.gov/bills/sb22-113 [https://perma.cc/7BNP-T7LL];
Connecticut Office of the Attorney General, *The Connecticut Data Privacy Act* (accessed Dec 1, 2024) https://portal.ct.gov/ag/sections/privacy/the-connecticut-data-privacy-act [https://perma.cc/2YN2-ECFH].

made and ensuring that human input is effectively solicited. In some cases, it may be clear that a particular AI system requires oversight at a particular point (e.g., the moment of deployment), but other important times for human intervention or participation can be overlooked (e.g., during early design or well after deployment). Adopting an inclusive view of human engagement from the start could encourage agencies to incorporate regular, structured points of reflection, outreach, and oversight that streamline AI adoption in the long run.

- **Create conditions that enable effective human oversight and engagement.** Human oversight encompasses a broad range of techniques, with some techniques more or less appropriate in different contexts. Existing research does not provide concrete answers about which techniques apply in a benefits delivery context. Techniques range from simple yes/no approval of an AI-recommended decision by a human to more complex frameworks where the human uses the AI in a more interactive way. Similarly, human engagement can produce various desired benefits, including making less harmful decisions, reducing bias in decisions, or imbuing determinations with qualitative considerations that are difficult to embed into an AI. Agencies should consider which approaches are best suited to their uses and monitor their processes over time to ensure they are imparting the desired benefits.
- **Allocate necessary resources (e.g., people, time, money) to ensure effective human oversight given the particular type and risk level of the AI use.** Agencies are often turning to AI systems to increase efficiency in managing their work. This means that the AI systems are intended to produce a large number of decisions and determinations (for instance, processing millions of benefits claims). Different human engagement frameworks may be more effective at different scales, requiring different types of expertise or training. In any case, these oversight systems will still require appropriate resources.
- **Create and use tools, interfaces, and frameworks to facilitate human oversight.** In order for humans to effectively oversee AI systems, they will need to intervene at the appropriate time and be given relevant information. This will necessitate tools that can provide the human with this data in a comprehensible and effective way.
- **Provide effective training and information for humans.** People who provide oversight will need to be trained to intercede in AI operations

and decisions. It is not currently clear what that training should look like and what competencies those involved will require, so agencies will need to monitor their chosen approach for effectiveness. In addition to competencies, these roles will need to be structured in ways that allow them to do their work effectively, with time and support to analyze and overturn AI decisions where necessary. The roles will also need ongoing support, monitoring, and training so that the humans will continue to challenge automation bias ("[an] overreliance on algorithmic advice even in the face of 'warning signals' from other sources"[5]). Agencies must design jobs with dedicated staff to encourage such oversight and training.

AI and algorithmic systems may have the potential to improve the public benefits landscape, but they also come with risks to the well-being of applicants and recipients that cannot be overlooked. Incorporating humans into the full lifecycle of algorithmic systems[6] — from the initial early planning stage on through development, procurement, deployment, and retirement — may help to improve the operation and positive impacts of these systems while limiting harms, but only if people are given the tools they need to meaningfully oversee these systems.

5 Saar Alon-Barkat & Madalina Busuioc, *Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice*, Journal of Public Administration Research and Theory (Feb 8, 2022) https://academic.oup.com/jpart/article/33/1/153/6524536 [https://perma.cc/9SZC-ZRYX].

6 Benefits Tech Advocacy Hub, *Understanding the Lifecycle of Benefits Technology* (accessed Sept 7, 2025) https://www.btah.org/lifecycle.html [https://perma.cc/X4R8-BHS5].

02 Introduction

Artificial Intelligence (AI) and other algorithmic systems (this report uses a broad definition of AI and algorithmic systems, see p. 13) are being deployed in all parts of society, including within agencies delivering public benefits. These tools are being incorporated into or even taking over tasks that were previously managed entirely by humans, such as determining the level of benefits a given person is eligible for (e.g., how many hours of home care or what level of income support they receive).

Generally, these systems are deployed with the stated goals of cutting program costs (e.g. lowering benefit levels or allocating scarce resources), reducing administrative burden, improving efficiency or customer service, preventing fraud and waste, improving program evaluation, and, in some cases, mitigating supposed inequities of human-only decision-making, like individual workers' biases. Although well-designed and appropriately governed AI systems can serve these goals, these systems sometimes produce outputs leading to agency decisions that cause real-world harm to the very people they exist to support. Such harm results when systems are deployed in contexts or for uses for which they are not appropriate or when they are not being effectively utilized, managed, and governed.[7]

7 Virginia Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*, St. Martin's Press (2018).

In cases where an algorithmic system may be appropriate, public agencies and advocates agree that one important guardrail for avoiding or mitigating the harms that can stem from such systems is to ensure adequate human oversight.[8] While not a panacea for algorithmic harms, human oversight improves system outcomes, limits biases that stem from the AI or algorithmic system, incorporates context that is either unavailable to or not comprehensible by the system, and, thus, reduces the likelihood of harm.[9]

8 Gina M. Raimondo & Laurie E. Locascio, *AI Risk Management Framework*, National Institute of Standards and Technology (Jan 2023) https://airc.nist.gov/airmf-resources/airmf/ [https://perma.cc/Q96M-39RU; https://perma.cc/NK96-SRQE]; City of Seattle, *Responsible Artificial Intelligence (AI) Program – Generative Artificial Intelligence Policy* (accessed Jun 16, 2025) https://www.seattle.gov/tech/data-privacy/the-citys-responsible-use-of-artificial-intelligence [https://perma.cc/5S8C-4V28; https://perma.cc/AC7A-QWR7];
Jeff Maxon, *Generative Artificial Intelligence Policy*, Kansas Office of Information Technology Services (July 25, 2023) https://www.governor.ks.gov/home/showpublisheddocument/405/638744386434630000 [https://web.archive.org/web/20250307073659/https://www.governor.ks.gov/home/showpublisheddocument/405/638744386434630000].

9 Ben Green, *The flaws of policies requiring human oversight of government algorithms*, Computer Law & Security Review (Apr 26, 2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921216 [https://perma.cc/XGN9-PMJK].

03 Defining Human Oversight

Researchers have considered and tested a range of human oversight and engagement systems. These approaches vary in one important way: whether they focus narrowly on human involvement solely at the point of an AI-based decision or whether they take a broader approach that contemplates more stakeholder involvement throughout the AI lifecycle, including whether AI is used in the first place. Some definitions focus solely on how much humans should be involved in a single given decision, whether that is by reviewing a complete AI output and making a keep-or-change decision, using output from the algorithm as a tool in their own decision-making, or providing input about the decision to the AI but leaving the AI to make the final determination.[10]

Other definitions of human oversight of AI focus on incorporating stakeholders throughout the AI lifecycle and in an AI system's overall design. Examples of this expanded definition include

10 Ben Green, *The flaws of policies requiring human oversight of government algorithms*, Computer Law & Security Review (Apr 26, 2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921216 [https://perma.cc/XGN9-PMJK].

consulting external stakeholders, involving a team in decisions rather than one individual, or monitoring systems using algorithmic auditing approaches.[11] As discussed in the best practices on p. 31, the latter approach is better suited to a public benefits context, ensuring that humans remain centered in these critical programs.

Although it is generally accepted that human oversight is important to any system that involves AI, it is not always clear what forms of oversight are needed for a given system in a particular context. Informed by existing scholarship, policy expertise, and experience challenging unjust uses of AI, this guidance aims to define the contours of human oversight of AI and provide recommendations for how to implement it if AI is incorporated into the administration of public benefits.

11 Deborah Morgan, Youmna Hashem, John Francis, Saba Esnaashari, Vincent J. Straub, & Jonathan Bright, *'Team-in-the-loop': Ostrom's IAD framework 'rules in use' to map and measure contextual impacts of AI* (Jun 30, 2024) https://arxiv.org/abs/2303.14007 [https://perma.cc/LGS4-B7DY];
Ben Green, *The flaws of policies requiring human oversight of government algorithms*, Computer Law & Security Review (Apr 26, 2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921216 [https://perma.cc/XGN9-PMJK];
Stuart E. Middleton, Emmanuel Letouzé, Ali Hossaini, & Adriane Chapman, *Trust, regulation, and human-in-the-loop AI within the European region*, Communications of the ACM (Mar 19, 2022) https://dl.acm.org/doi/10.1145/3511597 [https://perma.cc/52UW-VYG3];
Iyad Rahwan, (2018). *Society-in-the-loop: programming the algorithmic social contract*, Ethics and Information Technology (Aug 17, 2017) https://link.springer.com/article/10.1007/s10676-017-9430-8 [https://perma.cc/8KJG-A9NF];
Shea Brown, Jovana Davidovic, & Ali Hasan, *The algorithm audit: Scoring the algorithms that score us*, Big Data & Society (Jan 2021) https://www.semanticscholar.org/paper/The-algorithm-audit%3A-Scoring-the-algorithms-that-us-Brown-Davidović/ca79c5df0c966b959d39ac1fcba25b0c08c56149 [https://perma.cc/HR2Q-5736];
Brian Hedden, *On statistical criteria of algorithmic fairness,* Philosophy and Public Affairs (2021) https://philarchive.org/rec/HEDOSC [https://perma.cc/YA6F-P65M];
Briana Vecchione, Karen Levy, & Solon Barocas, *Algorithmic auditing and social justice: Lessons from the history of audit studies*, Equity and Access in Algorithms, Mechanisms, and Optimization (Oct 2021) https://dl.acm.org/doi/10.1145/3465416.3483294 [https://perma.cc/3FTU-EEQP].

04 AI and Algorithmic Systems in Public Benefits Delivery

As noted previously, AI can refer to a number of different types of systems deployed across the public benefits ecosystem for a range of stated uses.

AI and Algorithmic Systems

AI is a term with a wide range of often vague or contradictory definitions.[12] It can include complex machine learning systems that evolve over time as well as Large Language Models that power generative AI systems. In a public benefits context, this can include applications like chatbots that use generative AI to address user

12 Marko Grobelnik, Karine Perset, & Stuart Russell, *What is AI? Can you make a clear distinction between AI and non-AI systems?*, OECD (Mar 6, 2024) https://oecd.ai/en/wonk/definition [https://perma.cc/BDJ7-A9LD];
Hannah Quay-de la Vallee, Ridhi Shetty, & Elizabeth Laird, *Report – The federal government's power of the purse: Enacting procurement policies and practices to support responsible AI use*, Center for Democracy & Technology (Apr 29, 2024) https://cdt.org/insights/report-the-federal-governments-power-of-the-purse-enacting-procurement-policies-and-practices-to-support-responsible-ai-use/ [https://perma.cc/PQE8-3W2P].

queries or models that use AI to detect anomalous transactions.[13] At the same time, AI can serve as shorthand for any decision-making systems that use technology to apply a fixed set of rules (which may be large or small) and do not change over time unless manually adjusted by people. In terms of public benefits administration, examples of this more rudimentary technology can include data-matching technologies that compare information from multiple sources or rule-based assessments that assign "points" to applicants to determine what benefits they are eligible to receive.

When AI systems fail, they present problems distinct from those involved in failed human applications of rule-based criteria.

For purposes of this guidance, the term AI is inclusive of all of these applications of technology as they present overlapping and often similar potential as well as risks. However, it is important to note that this broad definition encompassing automated rule-based decision-making does not comport with all definitions of AI. Those who favor a narrower definition encompassing only the most sophisticated technologies suggest that less-sophisticated technologies that automate the application of eligibility rules feature a unique set of uses and concerns justifying separate treatment.

Nevertheless, this inclusive definition is used in this paper because of common risks that technology of varying levels of sophistication introduces. Such risks include inaccuracies when translating rule-based frameworks into code, inaccuracies or limitations in the input data (which may have been spotted by a person in a non-automated system), and the removal of people's ability to apply judgements to specific circumstances.[14] In practice, this has resulted

13 Kevin Levitt, *How Is AI used in fraud detection?*, NVIDIA (Dec 13, 2023) https://blogs.nvidia.com/blog/ai-fraud-detection-rapids-triton-tensorrt-nemo/ [https://perma.cc/DR9X-YUYN].

14 Human judgement can be an avenue to inject bias into a system, but it is also necessary to account for circumstances that fall outside the specific scenarios accounted for in the development of an automated system, which will never be able to account for all circumstances. Consequently, automation without human involvement is not an effective solution to bias. See, e.g. Ariana Aboulafia & Miranda Bogen, *To reduce disability bias in technology, start with disability data*, Center for Democracy & Technology (Jul 25, 2024) https://cdt.org/insights/report-to-reduce-disability-bias-in-technology-start-with-disability-data/ [https://perma.cc/3FR9-F878].

in repeated examples of failed implementation of automated rule-based systems for Medicaid, SNAP, and Unemployment Insurance.[15] When such systems fail, they present problems distinct from those involved in failed human applications of rule-based criteria: namely, larger scales of harm, more difficult detection of root causes, more obstacles to contesting adverse decisions, and more difficult individual and systemic fixes. These problems are similar enough to those raised by more narrowly-defined AI to warrant our broad approach.

While different systems may require different frameworks to ensure effective oversight, all of these sorts of systems may be operating in similar contexts and can cause similar harms. Additionally, those interacting with the system may not be aware that they are doing so, much less which type of system they are engaging with. Consequently, this brief will take a broad view of these sorts of systems and will use the terms algorithmic systems and AI interchangeably to highlight the diversity of systems that are relevant to this guidance.

Why do Agencies Use Algorithmic Systems?

Public benefits agencies adopt AI and algorithmic systems for a number of stated reasons, but there are a few major themes in agencies' stated goals in using AI. As discussed in more detail in the following section, the actual use of these systems often causes significant harm to people who apply for or receive benefits regardless of the stated goals.

15 Benefits Tech Advocacy Hub, *Case studies* (accessed Feb 4, 2025) https://www.btah.org/case-studies.html [https://perma.cc/N9GH-G7WY].

Cut Program Costs and Allocate Scarce Resources

AI systems can be used to allocate scarce resources or actually cut program costs. This may be in response to budget cuts, an increase in recipients or applicants without a corresponding increase in budget, or in response to political or program choices to target a program, with or without an explicit policy change.[16]

Reduce Administrative Tasks for Agency Staff

AI and algorithmic systems are often touted as a way to reduce the time spent on rote administrative tasks required of many agency personnel (e.g., data entry, mass correspondence), allowing the staff to spend time on work in which they specialize and that can have a greater impact on recipients. Systems in this category often include tasks like digitizing analog files, managing public inquiries, data input, and data transfer.[17] Robotic Process Automation, or RPA, (systems that "watch" human workers in order to automate repetitive tasks) is another common algorithmic tool for reducing administrative work.

16 Virginia Eubanks, *Want to cut welfare? There's an app for that.*, The Nation (May 27, 2015) https://www.thenation.com/article/archive/want-cut-welfare-theres-app/ [https://perma.cc/JPZ8-MPUU].

17 See, e.g. Emily Olsen, *Oracle to launch generative AI tools integrated with EHR*, Healthcare Dive (Sept 18, 2023) https://www.healthcaredive.com/news/oracle-health-generative-ai/693941/ [https://perma.cc/3QGA-NBDL]; Liza Lucas, *Georgia SNAP backlog may not be fixed until end of January, emails show*, 11Alive (Nov 16, 2023) https://www.11alive.com/article/news/local/georgia-snap-backlog/85-3b6fe6d3-62a3-460a-b0ce-cfb2c9ef20e7 [https://perma.cc/7AHL-A7BN].

Improve Customer Service

Algorithmic systems are also considered an avenue to improve customer service, allowing an agency to be more responsive and engaged, and improving people's ability to interact with the agency. One example of this is through the use of tools like chatbots, which are intended to serve as an always-accessible resource for applicants and recipients, providing help for people who are not always able to work within an agency's open hours.[18]

The primary concerns raised by these systems are effectiveness, accessibility, and accuracy. An ineffective system that does not meaningfully help applicants can waste their time, making applications more onerous for individuals who have insufficient time to meet all of their other obligations (e.g., work, caring for children or other family members, community commitments, household responsibilities). This may result in falloff in applications and loss of benefits for eligible people. Inaccessible systems, such as those that do not work with assistive technology or which are only available in English, may create a disproportionate barrier for disabled applicants and recipients or those with limited English proficiency. Finally, generative-AI systems that provide inaccurate or misleading answers to queries may result in failure to receive benefits or accusations of fraud for applicants and recipients who provide incorrect information or do not follow proper procedures because they have been misinformed.[19]

18 See, e.g. Louisiana Department of Health, *Medicaid offers virtual assist to members* (Nov 27, 2023) https://ldh.la.gov/news/7232 [https://perma.cc/HYH6-N8JP]; Arizona Health Care Cost Containment System, *Address change chatbot helps AHCCCS members update their contact information to prepare for renewal* (Jul. 13, 2023) https://www.azahcccs.gov/shared/News/PressRelease/AddressChangeChatbot.html [https://perma.cc/6UTK-MK55]; Dan Bateyko, *Let LLMs do the talking? Generative AI issues in government chatbots*, Center for Democracy & Technology (Dec 13, 2023) https://cdt.org/insights/let-llms-do-the-talking-generative-ai-issues-in-government-chatbots/ [https://perma.cc/5XKN-8ZYM]; Community Connect Labs, *Medicaid renewal: Digital strategy tips for member engagement* (Jun 15, 2023) https://communityconnectlabs.com/medicaid-renewal-tips-for-member-engagement-during-unwinding/ [https://perma.cc/PSL5-JDEP].

19 See, e.g. Colin Lecher, *NYC's AI chatbot tells businesses to break the law* (Mar 29, 2024) https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law [https://web.archive.org/web/20250123192245/https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law].

Prevent Fraud and Waste

Fraud by benefits recipients is rare and accounts for only a small amount of improper agency expenditures.[20] Still, agencies employ algorithmic systems to detect and reduce fraud and waste. AI approaches to fraud management typically include techniques like data analytics, identity verification, and document verification. However, there is significant concern about these systems' efficacy and fairness, and they have erroneously deemed legitimate recipients as fraudulent at immense scales, causing enormous financial, emotional, and social harm.[21]

20 Parker L. Gilkesson, *SNAP 'Program Integrity': How Racialized Fraud Provisions Criminalize Hunger, The Center for Law and Social Policy* (2022) https://www.clasp.org/wp-content/uploads/2022/04/2022_SNAP20Program20Integrity20-20How20Racialized20Fraud20Provisions20Criminalize20Hunger.pdf [https://perma.cc/JB33-PTUP];
Randy Alison Aussenberg, *Errors and Fraud in the Supplemental Nutrition Assistance Program (SNAP)*, Congressional Research Service (2018) https://sgp.fas.org/crs/misc/R45147.pdf [https://perma.cc/5QHM-RGUF].

21 Rachael Kohl, *Automated stategraft: Faulty programming and improper collections in Michigan's unemployment insurance program*, Wisconsin Law Review (Apr 22, 2024) https://wlr.law.wisc.edu/automated-stategraft-faulty-programming-and-improper-collections-in-michigans-unemployment-insurance-program/ [https://perma.cc/ZU7A-539U];
Ryan Felton, *Inside Michigan's faulty unemployment system that hit thousands with fraud*, The Guardian (Feb 12, 2016) https://www.theguardian.com/us-news/2016/feb/12/michigan-unemployment-insurance-benefit-automated-system-fraud-penalties [https://perma.cc/N4C2-2B8N];
Hannah Quay-de la Vallee, *Combatting identify fraud in government benefits programs: Government agencies tackling identity fraud should look to cybersecurity methods, avoid AI-driven approaches that can penalize real applicants*, Center for Democracy & Technology (Jan 7, 2022) https://cdt.org/insights/combatting-identify-fraud-in-government-benefits-programs-government-agencies-tackling-identity-fraud-should-look-to-cybersecurity-methods-avoid-ai-driven-approaches-that-can-penalize-real-applicant/ [https://perma.cc/2V9Q-WYQ4];
Shawn Donnan and Dina Bass, *How did ID.me get between you and your identity?* Bloomberg Businessweek (Jan 20, 2022) https://www.bloomberg.com/news/features/2022-01-20/cybersecurity-company-id-me-is-becoming-government-s-digital-gatekeeper [https://perma.cc/RT8Y-NQY6];
Lauren Hepler, *California's unemployment crash*, CalMatters (Nov 7, 2023) https://calmatters.org/series/california-unemployment-crash/ [https://perma.cc/5ZET-2U87];
Doron Dorfman, *Fear of the disability con: Perceptions of fraud and special rights discourse*, Law & Society Review (Oct 18, 2019) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3463814 [https://perma.cc/5TEE-V2EE];
Chelsea Barabas, Colin Doyle, JB Rubinovitz, & Karthik Dinakar, *Studying up: Reorienting the study of algorithmic fairness around issues of power*, Fairness, Accountability, and Transparency (Jan 27, 2020) https://dl.acm.org/doi/10.1145/3351095.3372859 [https://perma.cc/MB67-VQ4X].

When these systems fail or, in some cases, achieve their purpose of cutting costs by making benefits access more difficult, they can have dire consequences, such as erroneously stripping people of their healthcare, leading to long term health problems and even death.[22] Any resulting controversy can cost the agency time, public trust, and significant sums of money.[23]

Improve Program Evaluation

Public agencies also deploy AI and algorithmic systems in an effort to make their systems work better, making operations more timely, more consistent, and less error-prone. These uses can be aimed at program-wide evaluation and continuous improvement as well as at the individual recipient. For instance, AI-based systems can be used to analyze programs to identify inefficiencies or areas for improvement in program design or delivery, identify areas of high need, and flag issues like high drop-out or enrollment failure rates.[24]

22 Rachana Pradhan & Samantha Liss, *Medicaid for Millions in America Hinges on Deloitte-Run Systems Plagued by Errors*, KFF Health News (Jun 24, 2024) https://kffhealthnews.org/news/article/medicaid-deloitte-run-eligibility-systems-plagued-by-errors/ [https://perma.cc/XH8S-VB9F];
National Health Law Program, *National Health Law Program (NHeLP), EPIC, and Upturn Provide New Evidence to FTC in Deloitte Medicaid Eligibility Systems Complaint* Oct 16, 2024) https://healthlaw.org/news/national-health-law-program-nhelp-epic-and-upturn-provide-new-evidence-to-ftc-in-deloitte-medicaid-eligibility-systems-complaint/ [https://perma.cc/RP9C-6GXR];
Children's Defense Fund Texas, *In harm's way: True stories of uninsured Texas children* (Apr 2, 2007) https://www.childrensdefense.org/cdf-releases-new-report-in-harms-way-true-stories-of-uninsured-texas-children/ [https://perma.cc/8TBZ-ECBL; https://perma.cc/V7XK-QEES].

23 For example, litigation resulting from Michigan's decision to use a faulty AI system to detect fraud in its unemployment insurance program has resulted in $75 million of state liability. See, e.g., *UIA Fraud Class Action Settlement Website* (accessed Feb 4, 2025) https://uiaclassaction.com/ [https://perma.cc/TJ4X-YRPP];
Saunders v UIA Improper Collections Class Action (accessed Feb 4, 2025) https://bwclassactions.com/ [https://perma.cc/DY7Y-TQBV].

24 See, e.g. New York State Department of Health, *Request for Information on using advanced technology in Medicaid program integrity and efficiency* (Sept 21, 2020) https://www.health.ny.gov/funding/rfi/inactive/atpi/index.htm [https://perma.cc/QA5U-C83Q].

Improve Eligibility Determination for Benefits Applicants or Recipients

AI can be aimed at making individual benefits decisions more quickly, limiting application backlogs, and offering more standardization among benefits decisions. Such AI systems often involve the integration of multiple programs, such as Medicaid, the Supplemental Nutrition Assistance Program (SNAP), and Temporary Assistance to Needy Families (TANF). The AI system is supposed to take information from one program and use it to determine eligibility for others for which a person has applied. Due to the broad scope, AI system failures have extreme consequences: in multiple cases, failures in AI systems have led to improperly administered benefits programs.[25]

In select jurisdictions, AI systems have been used to indicate likely eligibility for a program for which a person has not applied.[26] Such information can be used by states to direct outreach to individuals or communities to encourage enrollment.

Also, in states where counties make eligibility decisions, AI systems may integrate and standardize the process across the many jurisdictions involved. The California Statewide Automated Welfare System (CalSAWS) is an example of such a system.[27]

25 See, e.g. National Health Law Program, Electronic Privacy Information Center, and Upturn, Inc., *Complaint and Request for Investigation, Injunction, and Other Relief in the Matter of Deloitte Consulting LLP*, (2024) https://healthlaw.org/wp-content/uploads/2024/01/NHeLP-EPIC-Upturn-FTC-Deloitte-Complaint.pdf [https://perma.cc/4Y2W-ECXX].

26 Digital Government Hub, *Hawai'i's coordinating SNAP & nutrition supports impact report* (2023) https://digitalgovernmenthub.org/library/hawaiis-coordinating-snap-nutrition-supports-impact-report/ [https://perma.cc/F3ZS-C47G].

27 California Department of Social Services, *California Statewide Automated Welfare System (CalSAWS)* (accessed Dec 1, 2024) https://www.cdss.ca.gov/inforesources/saws [https://perma.cc/Q3VM-R2ST].

How Algorithmic Systems Can Go Wrong

As shown previously, AI and algorithmic systems are often touted as a solution to the problems or challenges faced by benefits agencies. In practice, AI systems have often caused significant harm to the people these benefit programs are meant to serve. Among other harms, they have (a) inappropriately "rationalized" cutting or limiting of benefits; (b) inappropriately terminated eligible people; (c) resulted in false accusations of wrongdoing such as fraud, causing emotional and financial harms that can take years to recover from; (d) obscured decision-making frameworks, making it more challenging for recipients to argue their cases; (e) increased the administrative burden on recipients and applicants; (f) embedded and exacerbated bias and discrimination in benefits systems; (g) caused delays in the delivery of benefits in times of desperate need; and (h) limited recipients' and applicants' access to case workers, who often serve as key supports.[28]

Among this wide range of harms, there are a number of themes that emerge.

- **Bias in AI makes it harder for certain groups to access benefits.** AI systems frequently exhibit biases that can negatively impact people who interact with the system. In the benefits context, an example is facial-recognition-based identity verification systems. Many such systems have been shown to be less effective for people with darker skin tones, resulting in increased denials and a far more onerous verification process for those individuals.[29] Biases like these can lead to systems that are

28 Kevin De Liban, *Inescapable AI: The ways AI decides how low-income people work, live, learn, and survive*, TechTonic Justice (Nov 19, 2024) https://www.techtonicjustice.org/reports/inescapable-ai [https://perma.cc/8D53-UV6D].

29 National Employment Law Project, *ID Verification* (Nov 14, 2023) https://www.nelp.org/insights-research/id-verification/ [https://perma.cc/7D6C-MWYJ]; Benjamin Freed, *States warned about facial recognition for unemployment claims*, StateScoop (Apr 6, 2023) https://statescoop.com/labor-dept-inspector-general-warns-states-facial-recognition-unemploymen/ [https://perma.cc/TD78-ZHAP].

less effective or accessible for certain populations, particularly those already facing marginalization, for whom benefits are particularly critical.[30]

- **Privacy risks jeopardize trust and create a chilling effect on applying for benefits.** AI-based systems raise privacy concerns as they often require access to significant amounts of information for both training and day-to-day functioning. If the system is not engineered to protect this data, it can be another attack surface for hackers seeking to access sensitive input or training data or create additional privacy risks when linking previously separate data sets.[31] Additionally, generative AI systems, which produce novel content, could end up revealing sensitive information as part of a public output, such as through a chatbot system.[32] Moreover, AI perpetuates the structures of intensive surveillance that people receiving benefits face. The agglomeration of massive amounts of data from multiple sources, some of which may be inaccurate, threatens access to benefits and heightens risk of other unwarranted government intrusions, such as child welfare investigations and immigration enforcement. [33]

30 Mia Sato, *The pandemic is testing the limits of face recognition*, MIT Technology Review (Sept 28, 2021) https://www.technologyreview.com/2021/09/28/1036279/pandemic-unemployment-government-face-recognition/ [https://perma.cc/R5NE-4EUD].

31 Caitlin Chin-Rothmann, *Protecting data privacy as a baseline for responsible AI*, Center for Strategic & International Studies (July 18, 2024) https://www.csis.org/analysis/protecting-data-privacy-baseline-responsible-ai [https://perma.cc/3LYD-XWPX].

32 Ina Fried, *Generative AI's privacy problem*, Axios (Mar 14, 2024) https://www.axios.com/2024/03/14/generative-Ai-privacy-problem-chatgpt-openai [https://web.archive.org/web/20240424164918/https://www.axios.com/2024/03/14/generative-ai-privacy-problem-chatgpt-openai];
Jordan Pearson, *ChatGPT can reveal personal information from real people, Google researchers show*, Vice (Nov 29, 2023) https://www.vice.com/en/article/chatgpt-can-reveal-personal-information-from-real-people-google-researchers-show/ [https://perma.cc/Q2DM-37ET];
Roberto Torres, *Data privacy concerns swirl around generative AI adoption*, CIODive, (Sept 24, 2024) https://www.ciodive.com/news/deloitte-generative-AI-survey/727792/ [https://perma.cc/G7CV-2TLB].

33 Danielle Keats Citron, *A poor mother's right to privacy: A review*, Boston University Law Review (Jan 2018) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3100513 [https://perma.cc/8AHG-ARDY].

- **Ineffective and inaccurate systems divert resources from well-established practices.** In addition to wrongfully denying benefits to eligible people or intensifying discrimination, ineffective systems can eat up resources that could be better used elsewhere in the benefits delivery ecosystem.[34] Chatbots offer an example here: these bots are intended to increase the efficiency of a system and reduce load on agency employees. However, if the bot is not effective in this role — whether because it is not answering the right questions, is unable to provide answers specific to people's queries about their cases, or is providing fabricated information[35] — that money would likely be better spent on increasing staffing levels at the agency.[36] Harm from chatbots is exacerbated when it leads to decreased staffing, making it harder for people to access help to overcome the problems generated by such systems.
- **AI is ill-suited to measure certain concepts traditionally reserved to human discretion.** AI systems necessarily take in a restricted amount of information and are not adaptable to situations that do not conform to their model. So they are often ill-suited to handle cases that involve judgment, such as whether someone "needs" a particular service or "intends" to deceive the government to obtain benefits (fraud).[37] This can be particularly

34 Virginia Eubanks, *Want to cut welfare? There's an app for that.*, The Nation (May 27, 2015) https://www.thenation.com/article/archive/want-cut-welfare-theres-app/ [https://perma.cc/JPZ8-MPUU];
Arielle Dreher, *The 2017 Legislature's Lasting Effects on Mississippians*, Jackson Free Press (Apr 5, 2017) https://www.jacksonfreepress.com/news/2017/apr/05/2017-legislatures-lasting-effects-mississippians/ [https://perma.cc/W4QD-JBW6].

35 Lisa Lacy, *Hallucinations: Why AI makes stuff up, and what's being done about it*, CNET (Jul 1, 2024) https://www.cnet.com/tech/hallucinations-why-ai-makes-stuff-up-and-whats-being-done-about-it/ [https://web.archive.org/web/20240701182316/https://www.cnet.com/tech/hallucinations-why-ai-makes-stuff-up-and-whats-being-done-about-it/].

36 Sherin Shibu, *New York City's AI chatbot keeps getting facts wrong, 6 months and $600,000 after launch*, Entrepreneur (Apr 5, 2024) https://www.entrepreneur.com/business-news/nycs-first-ai-chatbot-keeps-getting-important-things-wrong/472280 [https://perma.cc/S739-WMNB].

37 Colin Lecher, *What happens when an algorithm cuts your health care*, The Verge (Mar 21, 2018) https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy [https://perma.cc/2UAZ-C4RL].

true for people with disabilities, who are often not well represented in data sets used to build AI and other algorithmic systems.[38]

- **AI-based decision-making is difficult to understand and contest.** Often, it is difficult for a person applying for or receiving benefits to know that AI is being used to make a decision about their eligibility. Even if it is known, understanding the decision is difficult. Generally, the notices that are sent contain confusing, contradictory, or vague information about the basis of the decision. Clarifying the notice by contacting the agency can take multiple trips to an office or several hours-long phone calls where no human contact is assured. Explanations for a denial or reduction in benefits will often be incomplete such that the person seeking benefits does not know what they must prove to get or keep benefits. Withholding such information generally violates due process requirements guaranteed to applicants or recipients (see at p. 26). Sometimes, agencies refuse to release information about the AI system on grounds that it is protected intellectual property or otherwise exempt from public disclosure. Beyond this, the statistical modelling that underlies the AI system is likely to be impenetrable without the aid of unaffordable technical experts. Even when accessible, such experts may not have the necessary time or resources available to study the system on the timeline required. Moreover, some systems may be so opaque that even technical experts are unable to assess them effectively. Ultimately, then, challenging the system as arbitrary or irrational is unlikely when recipients and advocates often lack sufficient information about the systems.

38 Ariana Aboulafia & Miranda Bogen, *To reduce disability bias in technology, start with disability data*, Center for Democracy & Technology (Jul 25, 2024) https://cdt.org/insights/report-to-reduce-disability-bias-in-technology-start-with-disability-data/ [https://perma.cc/3FR9-F878].

05 Current Status of Human Engagement and Oversight

In the face of these challenges and failures stemming from AI and algorithm systems, one suggested avenue for improving outcomes is to have human oversight of and engagement with algorithmic systems. While keeping "humans in the loop" is one such approach, this limited approach to human oversight relegates humans to a limited role in what should be a fundamentally human-centered system. Accordingly, this guidance calls for more robust human oversight. However, human oversight approaches vary in how inclusive they are, whether they are driven by existing legal requirements, and if research has shown them to be effective in mitigating risk.

Legal Requirements to Include Human Oversight in AI-Informed Systems

In some cases, there are legal requirements for human oversight of algorithmic systems. For instance, in Colorado, agencies that use a facial recognition system must ensure "that decisions that produce legal effects concerning individuals or similarly significant effects concerning individuals... are subject to meaningful human review."[39] Connecticut's consumer privacy law provides consumers with the

39 Colorado General Assembly, *SB22-113 Artificial intelligence facial recognition* (2022) https://leg.colorado.gov/bills/sb22-113 [https://perma.cc/7BNP-T7LL].

right not to be subject to a decision based solely on an automated process when this decision will produce legal effects for the consumer.[40] Though an opt-out right is not directly a mandate for human oversight, it is a framework that requires a human to be able to take control of any algorithmic decision system.

Benefits applicants or recipients face significant barriers to meaningfully challenge such systems through whatever recourse is available.

In addition, constitutional or statutory provisions requiring due process guarantee at least some limited human oversight when AI or algorithmic systems are involved in making a decision about an individual's eligibility for benefits. Generally, the person subject to the decision is entitled to notice of the decision, an explanation for the reasons it was made, and an opportunity to contest the decision in an administrative hearing over which a human hearing officer presides.[41] In practice, though, governmental bodies often do not fulfill due process mandates when AI or algorithmic systems are involved.[42]

Benefits applicants or recipients face significant barriers to meaningfully challenge such systems through whatever recourse is available. Among other obstacles, they often lack legal representation, cannot understand the opaque systems under challenge, and cannot easily navigate formal hearing procedures required to present evidence of eligibility. Additionally, decision-makers like administrative hearing officers often defer to the AI or

40 Connecticut Office of the Attorney General, *The Connecticut Data Privacy Act* (accessed Dec 1, 2024) https://portal.ct.gov/ag/sections/privacy/the-connecticut-data-privacy-act [https://perma.cc/2YN2-ECFH].

41 Jane Perkins, National Health Law Program, *Demanding Ascertainable Standards: Medicaid as a Case Study* (Mar. 2016), https://healthlaw.org/resource/demanding-ascertainable-standards-medicaid-as-a-case-study/ [https://perma.cc/954R-9BA9].

42 *See, e.g., Elder v. Gillespie*, 54 F.4th 1055 (8th Cir. 2022), *K.W. v. Armstrong*, 789 F.3d 962, 970-74 (9th Cir. 2015), *L.S. by & through Ron S. v. Delia*, No. 5:11-CV-354-FL (E.D.N.C. Mar. 29, 2012), *Jacobs v. Gillespie*, No. 3:16-cv-119-DPM (E.D. Ark. Nov. 1, 2016), *M.A. v. Norwood*, 133 F. Supp. 3d 1093, 1100 (N.D. Ill. 2015). Importantly, agencies often fail to fulfill due process mandates even when AI is not involved. AI exacerbates these baseline challenges.

algorithmic system's decision or lack the means to arrive at different decisions because AI has displaced other relevant criteria.[43]

Limitations of Human Oversight Approaches

Prior research on existing human oversight approaches of AI systems shows that oversight processes should be carefully tailored to the context and system in which they are implemented, to ensure that the particular approach is effective for the type of AI being used and the sorts of decisions to be made. Even so, they remain limited in the assurances they can provide and should be used alongside other AI governance approaches.[44] Research shows that human decision-makers using an algorithm as a supporting tool struggle to figure out how to interpret, judge the accuracy of, and incorporate scores or advice from algorithms effectively, even with training.[45] A human considering algorithmic information does not necessarily improve outcomes compared to making the decision alone. Even in instances when considering algorithmic information improves

43 Kevin De Liban, *Inescapable AI: The ways AI decides how low-income people work, live, learn, and survive*, TechTonic Justice (Nov 19, 2024) https://www.techtonicjustice.org/reports/inescapable-ai [https://perma.cc/8D53-UV6D];
Susan Landau, James X. Dempsey, Ece Kamar, & Steven M. Bellovin, *Challenging the machine: Contestability in government AI systems*, Workshop on Advanced Automated Systems, Contestability, and the Law (Jun 2024) https://arxiv.org/pdf/2406.10430 [https://perma.cc/L79V-UAVF].

44 Ben Green, *The flaws of policies requiring human oversight of government algorithms*, Computer Law & Security Review (Apr 26, 2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921216 [https://perma.cc/XGN9-PMJK].

45 Ujué Agudo, Karlos G. Liberal, Miren Arrese, & Helena Matute, *The impact of AI errors in a human-in-the-loop process*, Cognitive Research: Principles and Implications (2024) https://link.springer.com/article/10.1186/s41235-023-00529-3 [https://perma.cc/QB3S-S9YV];
Ben Green & Yiling Chen, *The principles and limits of algorithm-in-the-loop decision making*, Proceedings of the ACM on Human-Computer Interaction (Nov 7, 2019) https://dl.acm.org/doi/pdf/10.1145/3359152 [https://perma.cc/JM33-JQC8].

consistency of decisions, it often introduces other errors or biases.[46]

Humans can also exhibit their own biases when engaging with algorithms or use algorithm's outputs in biased ways. They have sometimes failed to act because a system failed to alert them to an issue, and they also often follow bad algorithmic advice even when there is evidence indicating it is erroneous.[47] One particular finding shows people tend to adhere to algorithmic advice more strongly when it aligns with stereotypes or existing racial biases.[48]

Given these shortcomings, there are a number of best practices, starting with expansive stakeholder inclusion, that can help to maximize the impact of human engagement in improving outcomes from algorithmic systems.

Who Are the Humans Who Need to Be Incorporated Into AI Oversight Practices?

Agencies using AI must ensure that the full range of people who will be impacted by an AI system or who have insights into its use and the related context are involved in system oversight. At

46 Ben Green, *The flaws of policies requiring human oversight of government algorithms*, Computer Law & Security Review (Apr 26, 2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921216 [https://perma.cc/XGN9-PMJK]; Ujué Agudo, Karlos G. Liberal, Miren Arrese, & Helena Matute, *The impact of AI errors in a human-in-the-loop process*, Cognitive Research: Principles and Implications (2024) https://link.springer.com/article/10.1186/s41235-023-00529-3 [https://perma.cc/QB3S-S9YV].

47 Ben Green & Yiling Chen, *The principles and limits of algorithm-in-the-loop decision making*, Proceedings of the ACM on Human-Computer Interaction (Nov 7, 2019) https://dl.acm.org/doi/pdf/10.1145/3359152 [https://perma.cc/JM33-JQC8].

48 Saar Alon-Barkat & Madalina Busuioc, *Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice*, Journal of Public Administration Research and Theory (Feb 8, 2022) https://academic.oup.com/jpart/article/33/1/153/6524536 [https://perma.cc/9SZC-ZRYX].

least three key groups are necessary: people seeking or receiving benefits and their advocates, agency staff using AI, and the people who build and deploy these systems. Agencies must account for differing levels of agency and power when incorporating these groups into oversight measures.

Agencies using AI must ensure that the full range of people who will be impacted by an AI system or who have insights into its use and the related context are involved in system oversight.

First, people whose data is fed into the AI system and who are subject to its decisions should have a right to participate in its use and oversight. They can offer insight into the ways the systems fail to operate equitably or effectively. As a starting point, applicants and recipients often understand through personal experiences the overall efficacy of public benefits programs: the paperwork involved and its associated burdens, the wait times required in offices or on call-in lines, and other difficulties in accessing or keeping benefits. Moreover, they often have different levels of familiarity and comfort with technology, providing a broad base from which to offer salient insights into AI systems' design, interfaces, ease of use, and, potentially, errors. Their allies or advocates — family members, caregivers, legal aid attorneys, and community support organizations, among others — may offer individual-specific information, spot broad trends, and recommend solutions to improve use or resolve inequities. Yet, despite this clear value, agencies do not generally offer formal public avenues for AI oversight apart from rulemaking procedures in the limited circumstances such procedures are required. Even then, rulemaking occurs only after the agency has finalized its plans, thereby limiting the likelihood that the agency will make significant changes in response to feedback. The only other available avenues for applicants, recipients, and their allies or advocates include administrative hearings, the limitations of which are discussed above, or advocacy requiring significant resources.

Second, staff using decisions or information from the AI must be involved because they are in a position to specify the purposes for which AI is used, guide its procurement and design, inform parameters for its use, use it to make decisions about benefit eligibility, train other staff, and incorporate experiences doing so to spot and correct errors or other inequities when in use.

Within agencies there are different functions of human oversight of AI systems that correspond to different types and levels of influence over the system. Effective human oversight of AI should incorporate them all, with particular regard for those agency staff best positioned to identify problematic uses or applications of AI to decisions about benefits eligibility. Importantly, such staff are likely to be caseworkers or other ground-level staff who directly use AI to make decisions but otherwise have little authority over the agency's big-picture practices. Agencies should thus recalibrate internal processes to properly solicit and credit input from these staff members.

Third, effective oversight requires input of the developers who design and build the AI system that will be used by the agency. Developers typically have expertise in AI or software development, but may not be experts in benefits policy or delivery. They generally have significant influence over AI system design and use through ongoing consultations with agency staff.

06

Guidance and Best Practices for a Human-Centered AI Approach

Emerging best practices, including those from well-established data governance and ethics practices, can support centering humans in systems that incorporate AI and thereby minimize the dangers:

- **Take an inclusive view when determining how and when people should be engaged throughout the lifecycle of a system**
- **Create conditions that enable effective human oversight and engagement**
- **Allocate necessary resources (e.g., people, time, money) to foster meaningful human oversight given the particular type and risk-level of the AI use**
- **Create and use tools, interfaces, and frameworks to facilitate human oversight**
- **Provide effective training, information, and quality control for humans**

Take an inclusive view when determining how and when people should be engaged

Humans should be involved both in decisions made by AI and decisions made about AI from inception to ongoing use. In particular, these decisions need input from the full range of people likely to be impacted by an AI system and its oversight framework. This means understanding when impactful decisions are being made and ensuring that human input is effectively solicited. Doing so may include actions like providing numerous engagement opportunities that are accessible to stakeholders with different scheduling, transportation, communication, language, or disability access needs. Oversight processes should be standardized by the agency to apply to all AI systems, with particular regard to those that make benefits decisions or otherwise interact with the public.

Recommendations:

- **Take a broad view of when to incorporate human oversight.** Human input should be solicited throughout the lifecycle of an AI system,[49] including:
  - » The current functioning of the benefits program under consideration, including how applicants or recipients want to the programs to work;
  - » Problem analysis and solution brainstorming, when the agency is thinking about what purposes AI might serve and what features or functions are needed from the AI;

49 Benefits Tech Advocacy Hub, *Understanding the lifecycle of benefits technology* (accessed Feb 4, 2025) https://www.btah.org/lifecycle.html [https://perma.cc/MWR9-WPD6].

- » Procurement, when the agency is seeking a vendor to build the AI that the government wants and seeks proposals from multiple companies in a document called a Request for Proposals (unless the agency chooses to build their system in-house);
- » Building, testing, and evaluation, when the agency is working with the chosen vendor to build the actual AI system that will be used, testing the system, and training staff;
- » The period prior to implementation, when the product is finalized but before final and formal approval; and
- » Deployment and ongoing monitoring, when the agency is actively using the AI system to inform or make decisions about recipients' benefits and assessing the impact of the system on recipients and agency operations.

Each of these steps should include input from a full array of agency staff including policy, technology, and end-user staff as well as the people who will be subjected to the algorithm. The input solicited should include a discussion of whether AI is an appropriate tool for the problem at hand. This discussion should be incorporated into ongoing monitoring of the system and should include feedback from people impacted.[50]

- **Take an inclusive view of which humans should be involved.** Every human that interacts with or is impacted by an AI system will possess information about the system's operation that might be critical to oversight and governance. Human oversight is often understood to mean those who manage the system such as agency employees who use it in the course of their work, but those who interact with the system in other ways can often provide input that the agency does not otherwise have. For instance, in a system that uses AI to try to match recipient records across different databases, recipients are better positioned than agency staff to flag errors, as they will know

50 London Borough of Camden, *Developing the Data Charter* (accessed Dec 1, 2024) https://www.camden.gov.uk/developing-the-data-charter [https://web.archive.org/web/20240526083208/https://www.camden.gov.uk/developing-the-data-charter].

when a record is not their own and may have prior experience with data matching issues. Therefore, the following perspectives should be considered and often involved in decision-making throughout the AI lifecycle: agency leadership, agency staff who will use the AI if deployed, agency technical staff who will maintain the system and monitor its performance, applicants and recipients about whom the system makes decisions or who interact with the system, and any other relevant stakeholders such as legal aid groups or community advocates. Direct consultation with applicants, recipients, and their allies can be informed and supplemented by information from appeals and customer service calls.

- **Ensure recipients and applicants have meaningful involvement.** Recipients and applicants or any oversight body they can join should have formal powers. If the formal powers do not include the ability to approve or deny AI uses, the body should have the powers to obtain information from the government agencies and contractors involved in developing and using the AI. This is needed because government agencies and contractors generally are not required to answer any questions and the documents they create may not be informative or accessible via public records laws like the Freedom of Information Act. This may also require translation or interpretation of highly technical documents, support in understanding them, and time needed to analyze them. For example, a standard 30-day comment period on a system that has taken years to develop may be insufficient due to the data used, policy assumptions being made, business rules, and other factors. There may also need to be staff support to help explain technical documents and provide additional information where needed.
- **Report to the public in an ongoing way.** Use of AI should not be a predetermined outcome or a one-time decision. Rather, there should be ongoing consideration of whether its use is appropriate and, if so, how the system is working in practice and what aspects need to be reconsidered or changed. Presently, the public has limited insight into an agency's decision-making processes, including how it vets, validates, tests, and monitors AI in development and in actual use. Accordingly, the agency should be required to regularly publish meaningful,

understandable information about the AI system and its impact on the communities subject to its decision. The reports should serve as periodic opportunities for the agency, oversight bodies, and authorities to decide with stakeholder input if the AI system is still justified in light of the harms it could cause or has caused, the risk of further harm, and the actual benefit to the agency from its use.

Create conditions that enable effective human oversight and engagement

Human oversight encompasses a broad range of techniques, and, as noted earlier, approaches range from simple yes/no approval of an AI decision by a human to more complex frameworks where the human uses the AI in a much more interactive way. Similarly, there are a number of different desired benefits of human engagement that include making more appropriate decisions, reducing bias in decisions, and imbuing determinations with qualitative considerations that are difficult to embed into an AI. Agencies should consider which approaches are best suited to their uses.

Recommendations:

- **Maintain transparent, comprehensive public-facing information about how AI is being used or seriously considered.** Comprehensive public-facing information, including comprehensive AI inventory[51] and context about why a given system was selected and how it performs is a key precondition for effective human oversight and engagement. This tool allows the agency to ensure that all AI systems they use have

51 Quinn Anex-Ries, *Best Practices for Public Sector AI Use Case Inventories*, Center for Democracy & Technology (Jul 21, 2025) https://cdt.org/insights/best-practices-for-public-sector-ai-use-case-inventories/ [https://web.archive.org/web/20250721165041/https://cdt.org/insights/best-practices-for-public-sector-ai-use-case-inventories/].

appropriate oversight and allows recipients, applicants, and the public to understand how AI is used in their case — including the specific factors and scores the AI uses — and how they may wish to engage. The inventory should be accessible to the community, including things like plain-language descriptions of the systems and any known limits, problems, risks, or biases. This inventory should be maintained in all languages spoken by the community.

- **Clarify the roles of algorithmic systems and agency staff involved in AI oversight.** For individual decisions or outputs, make the role of the algorithm and involved agency staff clear, including who bears responsibility for the decision and the level of control agency staff may exercise over the decision.[52] For instance, in a determination of level of care for health care services, is the involved agency staff allowed to entirely overrule the decision and assign a new level of care or are they only able to adjust the final decision by a certain margin? In cases where the agency staff does not have full latitude or has to justify deviations, document these frameworks clearly. Additionally, ensure the involved agency staff is able to fully exercise their control, meaning they have sufficient time and information to evaluate and adjust decisions. As noted in this report, agencies should track when and how agency staff intercedes, along with all necessary data to identify any inequitable uses of this power. All information about when, why, and for whom adjustments are made should be to improve the system as a whole.

- **Dedicate resources to support effective community engagement.** Ensuring that recipients and applicants who will be impacted by AI systems are able to engage involves building effective channels to gather their input throughout the AI system

52 Tor Grønsund & Margunn Aanestad, *Augmenting the algorithm: Emerging human-in-the-loop work configurations*, The Journal of Strategic Information Systems (Jun 2020) https://www.sciencedirect.com/science/article/pii/S0963868720300226 [https://perma.cc/M36M-J2WZ];
Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, et al., *Meaningful human control: actionable properties for AI system development*, AI and Ethics (May 18, 2022) https://link.springer.com/article/10.1007/s43681-022-00167-3 [https://perma.cc/8DMN-F5FF];
Lisanne Bainbridge, *Ironies of automation*, Automatica (May 23, 1983) https://www.sciencedirect.com/science/article/abs/pii/0005109883900468 [https://perma.cc/2WHA-J63F].

lifecycle (designing, building, and deploying those systems) and building capacity within the community.[53] Engagement should be solicited in a range of venues to maximize the range of people who can participate: paper or email surveys, town halls held in-person and remotely and scheduled at varied hours (including in the evenings when people working during the day can attend and child care may be more readily available), direct observation (such as watching applicants use systems like chatbots), and advisory councils. Community engagement should be conducted in all languages spoken by the community and in a fully accessible manner to meet the needs of community members with disabilities. For AI systems, agencies may need to build technical capacity amongst communities to allow them to better assess and evaluate the system. Those participating in community engagement processes should receive financial support for costs associated with participating, such as child care and travel. More broadly, participants should receive fair compensation for their oversight duties and that compensation should not be counted against public benefit financial eligibility standards. Engagement opportunities must include disability and language access measures so that disabled people and people who primarily speak languages other than English can fully participate.

- **Align incentives of involved agency staff to effectively manage AI-based decisions.** Make sure the incentives of agency staff with different relationships to the AI — for example, those who decide on its big-picture adoption and those who use it to make individual decisions — are clear and that they actually incentivize factors that lead to desired outcomes (e.g.,

53 Elizabeth Laird & Hugh Grant-Chapman, *Report – Sharing student data across public sectors: Importance of community engagement to support responsible and equitable use*, Center for Democracy & Technology (Dec 2, 2021) https://cdt.org/insights/report-sharing-student-data-across-public-sectors-importance-of-community-engagement-to-support-responsible-and-equitable-use/ [https://perma.cc/7BLA-7WN6].

not incentivizing speed of decisions).[54] Effective incentive alignment requires clearly defined and documented goals for AI and algorithmic systems. These documented goals are valuable for ensuring that agency staff are able to further those goals and measuring the impact of the system over time.

Allocate necessary resources (e.g., people, time, money) to ensure effective human oversight given the particular type and risk-level of the AI use

Often in response to resource constraints, agencies turn to AI systems in hopes of increasing efficiency in managing their work. This means that the AI systems are intended to produce a large number of decisions and determinations, such as processing millions of benefits claims. Different human engagement frameworks may be more effective at different scales, requiring different types of expertise, training, and oversight. Importantly, these oversight systems will require appropriate resources. Agencies should factor in resources required for oversight when weighing the benefits of adopting any AI system, as they will likely offset hoped-for efficiency gains.

Recommendations:

- **Monitor and oversee both individual decisions and the system as a whole.** In addition to overseeing the AI system's

54 Johann Laux, *Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act*, AI & Society (Oct 6, 2023) https://link.springer.com/article/10.1007/s00146-023-01777-z [https://perma.cc/B686-L39B];
Fabio Massimo Zanzotto, *Viewpoint: Human-in-the-loop artificial intelligence*, Journal of Artificial Intelligence Research (Feb 10, 2019) https://jair.org/index.php/jair/article/view/11345 [https://perma.cc/KAL6-45U9].

use in individual decisions, there should also be a framework to monitor, assess, and adapt the AI in systemic ways as needed, including assessing the systems' performance across demographics. Take the example of AI that is found to regularly underestimate the home-based care levels needed for Medicaid recipients with cerebral palsy. Rather than expecting front-line agency staff directly using the AI to adjust individual decisions for this population regularly, staff with greater authority should adjust the operation of the AI to correct this problem at the root. This oversight should extend to contextual factors as well. So, for instance, if the rate of overruling individual AI decisions by staff directly using the AI suddenly drops, higher-level agency staff should investigate and determine why this is happening and if process or training changes are needed. To do this effectively, the AI system needs to be appropriately assessed in context, agency staff directly using AI need avenues to provide feedback, and staff with greater authority need time and resources (including access to technical expertise independent of system vendors) to evaluate the AI system.[55]

- **Provide staff directly using AI systems with appropriate time to evaluate AI decisions.** Assess the appropriate workload for agency staff directly using the AI. As caseloads and algorithmic decisions increase, so should the staff numbers and staff time devoted to meaningful oversight. Agencies should be extremely cautious about adopting AI systems if they do not have the human capacity for long-term oversight. This includes capacity for community engagement with recipients and applicants, capacity to manage individual decisions and supervise the system overall, capacity to adjust the system as needed, and capacity to decommission the system if necessary. Agencies should also consider adopting a stance of presumed positive determinations if there are periods where the agency is unable to provide oversight of the system in a timely manner causing delays in determinations.

55 Ben Green, *The flaws of policies requiring human oversight of government algorithms*, Computer Law & Security Review (Apr 26, 2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921216 [https://perma.cc/XGN9-PMJK].

Create and use tools, interfaces, and frameworks to facilitate human oversight

In order for humans to effectively oversee and govern AI systems, they will need to intervene at the appropriate time and be given relevant information. This will necessitate tools that can provide the human with this data in a comprehensible and effective way.[56]

Recommendations:

- **Provide people with information from and about the AI in a contextually useful way.** During community engagement, ensure that people have the information they need to understand and evaluate the potential merits and risks of the system. Experiment with the order and format in which information is presented to agency staff directly using AI relative to when they make initial or final judgements. Whether they get information from the algorithm before making a preliminary judgment can impact how agency staff use information. For example, staff can be swayed by erroneous algorithmic decisions, but that impact is ameliorated if staff receive input from the AI after making their own judgement.[57] To maximize the effectiveness of combined

56 Lisanne Bainbridge, *Ironies of automation*, Automatica (May 23, 1983) https://www.sciencedirect.com/science/article/abs/pii/0005109883900468 [https://perma.cc/2WHA-J63F];
Barry Strauch, *Ironies of automation: Still unresolved after all these years*, IEEE Transactions on Human-Machine Systems (Aug 18, 2017) https://www.jurispro.com/files/articles/roniesofutomationtillnresolvedfterllheseears_4830.pdf [https://perma.cc/6P8H-VAHP];
Mary Missy Cummings, *Man versus machine or man + machine?*, IEEE Intelligent Systems (2014) https://www.computer.org/csdl/magazine/ex/2014/05/mex2014050062/13rRUILLkzS [https://perma.cc/C3TD-WZPR].

57 Ujué Agudo, Karlos G. Liberal, Miren Arrese, & Helena Matute, *The impact of AI errors in a human-in-the-loop process*, Cognitive Research: Principles and Implications (2024) https://link.springer.com/article/10.1186/s41235-023-00529-3 [https://perma.cc/QB3S-S9YV].

decision-making, build in flexible design patterns to enable in-time human corrections or curate additional explanations as needed.[58]

- **Provide agency staff with AI information in context with non-AI information.** Information and determinations from the AI should be presented alongside other relevant information so that AI-sourced information is not given undue weight. To the greatest extent possible, agency staff directly using AI should have access to the raw data that was input by a user, claimant, or recipient, not just the AI system's interpretation or analysis of that information.
- **Provide information about the AI decision itself.** The agency staff should be able to see things like what factors contributed to the AI's decision and how they were weighed, why certain weights were deemed appropriate, how confident the AI is in its determination (and what particular confidence levels mean), and limitations of the AI's process.[59]
- **Provide information about the AI's systemic performance.** Agency staff of all authority levels should be made aware of the AI system's level of performance across different metrics like accuracy and consistency across demographics, as well as how those metrics are defined and analyzed, so they can calibrate if and how to use and assess it.[60] Information and training should emphasize that the algorithm can be wrong and emphasize the value of staff discretion. Staff directly using AI systems should

58 Projects By If, *Design patterns catalogue* (accessed Dec 1, 2024) https://catalogue.projectsbyif.com/ [https://perma.cc/XKA2-MVRJ].

59 José J Cañas, *AI and ethics when human beings collaborate with AI agents*, Frontiers in Psychology (Mar 2022) https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.836650/full [https://perma.cc/ZDC3-UK8S]; Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, et al., *Meaningful human control: actionable properties for AI system development*, AI and Ethics (May 18, 2022) https://link.springer.com/article/10.1007/s43681-022-00167-3 [https://perma.cc/8DMN-F5FF].

60 Ben Green & Yiling Chen, *The principles and limits of algorithm-in-the-loop decision making*, Proceedings of the ACM on Human-Computer Interaction (Nov 7, 2019) https://dl.acm.org/doi/pdf/10.1145/3359152 [https://perma.cc/JM33-JQC8].

be kept informed about the function of the system, including known bugs or common failure cases, and should be able to give feedback on their experience using and overseeing the system. Aggregate system performance should be made public so that recipients, applicants, and the wider public can engage in shared decision-making and provide feedback on its use. This data should include information about how the system performs across different demographics to allow stakeholders to assess the system for biased outcomes and make corrections. Information about errors and AI system changes should be posted publicly to maximize stakeholders' awareness and be provided to applicants or recipients potentially affected by them, including by mail, application portals, and call center staff.

- **Smooth the transition from human-only to AI-assisted decision-making processes.** Software tools that facilitate human-AI decision-making should have easy-to-understand user interfaces and should help connect the structure of the new algorithm-and-human decision-making process to the process the involved agency staff previously used and is familiar with.[61] To the fullest extent possible, any transition to AI-assisted decision-making should also include procedural fail-safes: if an AI system stops working as intended or causes unforeseen harm, agency staff must be able to efficiently move to a human-only process.
- **Provide agency staff with meta-tools that enable them to review outputs and screen for systemic problems.** Software tools should be implemented to enable data dashboards for internal review that allow agency staff, particularly staff with significant authority, to monitor where AI-driven systems may be leading to unintended outcomes, such as large numbers of notices suddenly being generated for a particular issue. These meta-tools can feed back into system updates and help identify common issues that need to be addressed.

61 Lanthao Benedikt, Chaitanya Joshi, Louisa Nolan, Ruben Henstra-Hill, Luke Shaw, & Sharon Hook, *Human-in-the-loop AI in government: A case study*, Intelligent User Interfaces (Mar 17, 2020) https://dl.acm.org/doi/10.1145/3377325.3377489 [https://perma.cc/T7CY-VUMP].

Provide effective training and information for humans

Agency staff will need to be trained to intercede in AI decisions, and it is not currently clear what that training should look like and what competencies staff will need. There must also be quality control mechanisms to ensure that staff are interceding when needed. In addition to competencies and oversight, roles will need to be structured in ways that allow staff to do their work effectively, including time and support to analyze AI decisions and overturn them where necessary. Agencies must design jobs to encourage effective oversight and provide appropriate training.

Recommendations:

- **Provide clear goals and success metrics for the human engagement framework.** Training should make clear the goals of human oversight. This should include examples of times when an AI determination is found to be incorrect or undesirable or situations where AI-based information was found to be less important than other, countervailing data. Ensure that the goals and success metrics are aligned with the intended purpose of human oversight and do not create conflicting incentives, such as encouraging rapid review that does not allow time for meaningful analyses of decisions.
- **Provide accessible education for recipients and applicants.** For AI systems, agencies may need to build technical capacity amongst participant communities to allow them to better assess and evaluate the system. This education should provide information about how the AI works, how it is used, and what the human oversight process entails. The education should be built specifically for the community the agency serves, with due regard for the level of starting technical knowledge, languages spoken by the community, and cultural sensitivities.
- **Train staff about the AI decision-making process.** Training should include explanations of how the AI system works,

including how it makes decisions or calculates scores.[62] The level of explanation matters greatly (unnecessary explanations can be unhelpful or harmful), so agencies should iterate their trainings over time to experiment with what staff find most helpful.[63] To avoid criteria displacement, ensure that staff understand the limitations of the AI system, including the limitations of the data and context the system is able to incorporate into its decisions. This may include considerations like different presentation of a disability, different contextual factors in the life of the recipient, and what services the agency is able to provide. If engagement with the community has surfaced goals and concerns from the community, incorporate those perspectives into trainings.

- **Do not anthropomorphize AI systems.** Attributing human capabilities to AI systems can cause staff to over-value the algorithm's outputs. Therefore, training should avoid assigning personal agency or human capabilities to the AI systems and should explain to staff that they should avoid anthropomorphizing the system themselves.[64] For example, a statement like "the computer has a mind of its own" may lead some to assume the system is smarter than them, when it is actually just inscrutable.
- **Address AI bias concerns, including automation bias, in trainings.** AI systems should be designed to limit bias. However, even if they are, training programs should include modules on the ways that AI systems can exhibit bias and how to actively combat that bias. Training should emphasize that no AI will be bias- or error-free and that staff oversight is a critical tool in limiting impacts of AI bias. Conveying the complexities of AI bias requires acknowledging the biases in human judgment, decisions, and actions that are reflected in training data. The agency must also test and monitor the efficacy of training and human intercession.

62 Ben Green, *The flaws of policies requiring human oversight of government algorithms*, Computer Law & Security Review (Apr 26, 2022) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3921216 [https://perma.cc/XGN9-PMJK].

63 José J Cañas, *AI and ethics when human beings collaborate with AI agents*, Frontiers in Psychology (Mar 2022) https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.836650/full [https://perma.cc/ZDC3-UK8S].

64 Nanyi Bi and Janet Yi-Ching Huang, *I create, therefore I agree: Exploring the effect of AI anthropomorphism on human decision-making*, Computer Supported Cooperative Work and Social Computing (Oct 14, 2023) https://dl.acm.org/doi/abs/10.1145/3584931.3606990 [https://perma.cc/4JWD-DTVT].

07 Conclusion

AI and algorithmic systems may have the potential to improve the public benefits landscape, but they also come with grave risks to the well-being of applicants and recipients, as seen in the many real-world implementations resulting in erroneously denied benefits, false fraud accusations, or other similar harms.

Incorporating humans into the full lifecycle of algorithmic systems may help to improve the operation and positive impacts of these systems while limiting harms.

However, human involvement will only be a reliable component of accountability if the people are given the structures, tools, and resources they need to meaningfully engage with and oversee these systems.

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

BENEFITS
TECH
ADVOCACY
HUB

CENTER FOR
DEMOCRACY
& TECHNOLOGY