JOBS AND MASS UNEMPLOYMENT

Central question: What happens when companies no longer need as many human workers?

A. ISSUE OVERVIEW

- AI can write, summarize, code, and operate machines. Automating some tasks can shrink a workforce, increase output with the same staff, or create demand for different work; these outcomes should not be conflated.
- Work supplies income, health coverage, social standing, and a route into adulthood. Even if total employment eventually recovers, displaced workers can face years of lower earnings, while young people lose the entry-level jobs that build expertise.
- Owners and consumers may gain from lower costs, and workers may benefit from tools that make them more productive. Clerical workers and some professional workers face substantial exposure, but physical jobs depend on advances in robotics as well as language models; gender, savings, bargaining power, and geography affect vulnerability.
- U.S. employers are adopting AI while announcing layoffs, although company explanations rarely isolate AI's contribution. A large customer-support study found productivity gains, demonstrating that assistance can improve human performance. The ILO finds broad global exposure but expects transformation to be more common than complete replacement. Across countries, governments are weighing training and income protection, while adoption remains uneven and causal evidence on economy-wide employment is limited.
- The key unresolved question is whether new demand and new tasks will offset displacement quickly enough. A second is who captures the gains: workers with negotiating power, consumers facing competitive prices, or owners of the systems.

B. KEY FACTS AND EVIDENCE

- A study of 5,172 customer-support agents found that AI assistance increased issues resolved per hour by 15% on average. This is evidence from one workplace, not a national employment forecast. Customer-support study.
- The ILO estimates that one in four workers worldwide has some occupational exposure to generative AI; 3.3% are in its highest exposure category. Exposure measures potential task change, not layoffs. ILO exposure index.
- In the ILO's highest category, exposure is 4.7% of women's employment and 2.4% of men's globally, reflecting occupational differences rather than a prediction about individual workers. ILO gender estimates.
- Anthropic's January 2026 report classified 52% of sampled November 2025 Claude.ai conversations as augmentation and 45% as automation. These vendor-specific interaction labels do not measure jobs saved or lost. Economic primitives.
- Recent AP reporting finds companies invoking AI when cutting jobs while also citing restructuring and broader economic conditions. An employer's announcement alone cannot establish that AI caused the reductions. Layoff reporting.
- One published economic model estimated no more than a 0.66% increase in total factor productivity over ten years under its assumptions. That deliberately modest projection illustrates how uncertain aggregate forecasts remain; it is not a measured outcome. Acemoglu's model.

C. RESEARCH AND SOURCE LIBRARY

1. Peer-reviewed academic research

Generative AI at Work — Erik Brynjolfsson, Danielle Li, Lindsey Raymond; 2025; The Quarterly Journal of Economics. A staggered workplace rollout found uneven productivity gains, particularly for less-experienced staff. Useful evidence for assistance and training, with limited generalizability to other occupations or employment totals.

2. Government and intergovernmental reports

Generative AI and Jobs: A Refined Global Index of Occupational Exposure — Paweł Gmyrek and colleagues; ILO Working Paper 140; May 2025. Combines task assessments and labor data to map exposure. It supports targeted transition planning; it is an intergovernmental working paper, not a peer-reviewed estimate of future unemployment.

3. AI industry research

Anthropic Economic Index report: Economic primitives — Anthropic; January 2026. Analyzes anonymized Claude interactions and distinguishes assistance, autonomy, and task success. This is company research on its own users, with selection and model-classification limitations and a commercial interest in AI adoption.

4. Independent think tanks and nonprofit organizations

Workforce policy for the age of AI — Brookings; 2026. Proposes sector-specific training, apprenticeships, support for displaced workers, and wage insurance. Its concrete contribution is a transition framework that can be tested through pilots rather than assuming a single unemployment scenario.

5. Credible journalism

From Cisco to Block, more companies are pointing to AI when unveiling job cuts — AP; May 2026. Examines corporate explanations for layoffs. Shows why organizers should distinguish direct substitution from spending shifts and restructuring.

As the coding boom fades, computer science grads focus on AI skills in choppy job market — AP; September 2026. Reports graduates' hiring difficulties and schools' responses. Makes career-ladder disruption concrete without proving that AI explains all weak hiring.

Is AI causing tech worker layoffs? That's what CEOs suggest, but the reality is complicated — AP; 2025. Investigates hiring trends, employer claims, and alternative explanations. Useful preparation for questions about causation.

Exclusive: Anthropic's “index” tracks AI economy — Axios; February 2025. Covers the initial release of company usage data and the information advantage providers hold. Explains why independent access to labor-relevant evidence matters.

6. Strongest opposing perspective

The simple macroeconomics of AI — Daron Acemoglu; 2025; Economic Policy. Challenges sweeping economic forecasts using a task-based model and modest estimated aggregate gains. The disagreement turns on adoption, reliable task completion, new tasks, and demand responses; longitudinal employer and worker data would help resolve it.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Measure displacement before declaring it** [Near-term priority]. Congress should fund BLS and Census surveys linking AI adoption to hiring, hours, pay, and separations, with privacy-protected researcher access. Exposure measures in the ILO report justify monitoring, but cannot replace causal analysis; reporting should be sampled and proportionate for small firms.
2. **Extend transition notice** [Near-term priority]. Congress should amend the WARN Act to cover substantial AI-driven reductions in hours as well as qualifying layoffs, and require a plain explanation of affected tasks and transition support. Advance notice gives workers time to respond, but coverage thresholds and emergency exceptions should avoid discouraging small-firm hiring.
3. **Bargain over workplace AI** [Structural reform]. Congress should clarify bargaining duties for material AI changes affecting staffing, evaluation, and workload, backed by NLRB enforcement. Worker consultation can shape how productivity gains are shared; contracts should not give either side an unlimited veto over ordinary software upgrades.
4. **Pilot wage insurance** [Near-term priority]. Congress should authorize time-limited payments covering part of the earnings loss when displaced workers take lower-paid jobs, administered

through state workforce agencies. Brookings identifies this practical transition tool; caps, independent evaluation, and minimum job-quality standards are needed to limit employer wage subsidies.

5. **Modernize unemployment insurance** [Structural reform]. Congress should finance interoperable state systems and minimum access standards covering eligible part-time workers and unpredictable scheduling. Faster income replacement cushions disruption without needing to prove AI caused each layoff, although expanded eligibility requires stable funding and safeguards against wrongful denials and fraud.
6. **Fund paid, job-linked retraining** [Near-term priority]. Labor Department grants should pay tuition and living stipends for employer-linked community-college programs, with earnings and completion outcomes published by program. Assistance gains in the workplace study support learning relevant tools; grants should end when providers repeatedly fail to improve employment outcomes.
7. **Protect the entry-level ladder** [Near-term priority]. Congress should subsidize registered apprenticeships in exposed occupations, requiring paid mentorship, genuine skill development, and retention reporting. Recent graduate reporting supports attention to lost pathways; subsidies need additionality checks so firms cannot relabel existing hires as new opportunities.
8. **Share productivity gains in public contracts** [Near-term priority]. Federal procurement agencies should pilot contract incentives tied to verified worker training, job quality, or gain-sharing when AI reduces delivery costs. The productivity study supplies a reason to test shared gains; procurement officials must measure outcomes rather than reward unenforceable promises.
9. **Use work-sharing before layoffs** [Near-term priority]. Congress and states should expand short-time compensation, allowing unemployment insurance to replace part of wages when firms temporarily reduce hours across staff. This can preserve skills during restructuring; participation should require worker notice and avoid indefinitely supporting unviable businesses.
10. **Remove tax favoritism for substitution** [Structural reform]. Treasury and Congress should review tax provisions that subsidize capital while taxing payroll, then consider revenue-neutral changes based on demonstrated distortions. A blanket “robot tax” is hard to define and may suppress beneficial investment; evaluate targeted corrections and transition funding before taxing every automated task.
11. **Make benefits portable** [Structural reform]. Congress should establish enforceable contribution and vesting rules for a portable benefits system covering workers who move between employment and contract work. Mobility becomes safer when health and retirement support persist; design must prevent firms from using portability to evade employee status and existing obligations.
12. **Aid places facing concentrated losses** [Near-term priority]. Commerce and Labor should direct competitive adjustment grants to areas with sustained employment and earnings declines, using verified data rather than an AI-exposure score alone. Local infrastructure and business diversification can help communities; independent review should block politically favored projects without plausible jobs.
13. **Test a basic-income floor** [Long-term proposal]. Congress should fund rigorously evaluated guaranteed-income pilots with a preannounced budget, benefits coordination, and employment and well-being measures. Such pilots explore insurance against persistent displacement, but neither cost nor labor-market effects can be inferred from exposure forecasts or small voluntary experiments alone.
14. **Offer public jobs where demand fails** [Long-term proposal]. Congress should pilot a voluntary employment guarantee in severely affected regions, using local governments and nonprofits to provide useful care, conservation, and infrastructure work at an explicit wage. This supplies an option when private jobs are scarce; supervision, displacement of existing public staff, and sustained financing are major constraints.
15. **Pay for transition through broad revenues** [Structural reform]. Congress should create a standing worker-transition fund financed through general taxation and, where justified, stronger taxation of economic rents rather than a per-model levy. This links adaptation to shared prosperity without

needing to identify an “AI job”; revenue volatility and effects on investment require regular review. See Tab 2 for ownership and competition proposals.

E. STRONGEST COUNTERARGUMENTS

“Technology creates jobs; intervention will slow growth.” Historical job creation and measured assistance gains make this serious. They do not establish that the same workers recover quickly, so targeted insurance and paid training are more defensible than assuming either permanent mass unemployment or painless adjustment; the scale and timing of net effects remain uncertain.

“Workers need skills, not income programs.” Skills can improve mobility, but a person cannot train without rent money or suitable openings. Pair training with temporary support and evaluate earnings, recognizing that neither a certificate nor wage insurance guarantees a durable career.

“Automation taxes and job guarantees will waste money.” Defining automation is difficult, and public jobs require management and financing. Favor broad funding and evaluated regional pilots first; the case for national expansion depends on sustained displacement, measurable program benefits, and feasible costs.

F. AI ACTION ADVOCACY BRIEF

Core message: Workers deserve a share of AI's gains and a real path forward when their jobs change.

Three strongest talking points

- AI exposure is a measure of possible task change, not a count of future layoffs.
- A large workplace study found productivity gains from helping people do their jobs.
- Transition policy should protect earnings and career ladders, even when total employment holds up.

Three priority policy proposals

- Paid retraining tied to actual jobs (6).
- Wage-insurance pilots with outcome evaluation (4).
- Bargaining rights over material workplace AI changes (3).

Questions for lawmakers

- Will you support paid transition training with public earnings results?
- How will workers negotiate over AI-driven changes to staffing and workload?
- What evidence would trigger additional income support in a badly affected region?

What we still don't know

- How much AI changes net employment after demand and new tasks respond.
- Whether early-career opportunities recover or permanently narrow.
- Which transition programs deliver durable earnings gains at reasonable cost.

CONCENTRATION OF WEALTH AND CORPORATE POWER

Central question: Who owns the technology that could produce enormous amounts of wealth?

A. ISSUE OVERVIEW

- AI's economic gains depend on who owns chips, cloud infrastructure, models, applications, and rights to useful data. Concentration at one layer can give firms leverage over businesses at another, even when customers can choose among several chatbots.
- This is a competition and distribution problem, not simply a question of how rich technology executives become. Control over essential inputs can determine prices, research priorities, and whether smaller businesses can challenge established firms.
- Large investors, infrastructure suppliers, and successful developers can capture substantial returns. Consumers and small companies may benefit from cheaper services, but workers, creators, public researchers, and dependent app developers can lose bargaining power when alternatives are weak.
- A 2025 FTC study documented contractual ties between major cloud providers and leading developers, while stopping short of a definitive competition assessment. U.S. enforcement debates concern partnerships as well as acquisitions. Britain's competition authority concluded in 2025 that Microsoft's OpenAI arrangement did not qualify for the merger review it was considering. More recently, debate has expanded to model providers competing directly with the app businesses that buy their services, while cheaper and open models complicate assumptions about durable dominance.
- The central uncertainty is whether scale advantages will harden into exclusionary power or be undermined by efficiency improvements and new entrants. Another is how society can share gains without discouraging productive investment or handing officials discretion to favor politically connected firms.

B. KEY FACTS AND EVIDENCE

- The FTC's January 2025 study examined three partnerships involving five firms: Microsoft–OpenAI, Amazon–Anthropic, and Alphabet–Anthropic. This is a defined sample, not a census of the whole AI market. FTC staff report.
- Some studied agreements committed developers to spend substantial portions of investors' funding on the investors' cloud services. That can finance entry while also creating dependence. Partnership terms.
- The FTC identified potential switching costs and access to sensitive developer information as areas requiring scrutiny. It did not establish that every partnership was unlawful. Competition implications.
- Ahmed, Wahed, and Thompson documented industry dominance in leading models and research inputs. Historical evidence of concentration is a reason to examine access, not proof that today's market cannot be disrupted. Research on industry influence.
- DeepMind's Chinchilla experiment found a smaller model could outperform a much larger one using the same training compute budget and more training data. Parameter count alone is an unreliable guide to competitive barriers. Compute-efficiency research.
- The UK watchdog dropped its Microsoft–OpenAI merger review in March 2025 because the arrangement did not meet the relevant change-of-control test. That jurisdictional conclusion was not a blanket finding that all AI markets are competitive. UK review reporting.

C. RESEARCH AND SOURCE LIBRARY

1. Peer-reviewed academic research

The growing influence of industry in AI research — Nur Ahmed, Muntasir Wahed, Neil C. Thompson; 2023; Science; DOI 10.1126/science.ade2420. This research-based perspective maps industry's growing command of compute, data, talent, and research outputs. It supports public research capacity, while offering no direct estimate of future monopoly prices.

2. Government and intergovernmental reports

Partnerships Between Cloud Service Providers and AI Developers — FTC staff; January 2025. Uses compulsory information gathering to document investment, commercial, and information-sharing arrangements. Useful for identifying potentially restrictive terms; the report explicitly does not complete a competition-effects assessment.

3. AI industry research

An empirical analysis of compute-optimal large language model training — Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Laurent Sifre and colleagues; Google DeepMind; 2022. Technical experiments show how training choices can improve performance at a fixed compute budget. They challenge simplistic scale arguments; the company has interests in both model development and cloud competition.

4. Independent think tanks and nonprofit organizations

What happens when AI companies compete with their customers? — Mark MacCarthy; Brookings; March 2026. Examines conflicts when model suppliers enter downstream app markets. Proposes limits on unreasonable discrimination, an actionable alternative to treating every large provider as an automatic monopoly.

5. Credible journalism

Nvidia overcomes tariff-driven turbulence to deliver Q1 results that eclipsed projections — AP; May 2025. Documents demand and financial gains at a key chip supplier. Illustrates where revenue accumulates without establishing an antitrust violation.

UK watchdog drops competition review of Microsoft's OpenAI deal — AP; March 2025. Explains the regulator's change-of-control conclusion. Useful evidence about the limits of merger jurisdiction.

Nvidia CEO Jensen Huang visits Republicans as debate over intensifying AI race rages — AP; December 2025. Reports lobbying and policymaker access around chip restrictions. Makes corporate political influence concrete without assuming that access proves capture.

OpenAI may move forward with new business structure, partnership with Microsoft, regulators say — AP; October 2025. Examines nonprofit control, investor interests, and state oversight of restructuring. Shows why ownership and governance must be assessed separately.

6. Strongest opposing perspective

Ferguson and Holyoak's statement on the AI partnerships study — FTC commissioners Andrew N. Ferguson and Melissa Holyoak; January 2025. Supports vigilance against incumbent control but argues the narrow study cannot justify sweeping predictions or premature regulation. Evidence on entry, switching, prices, and exclusion would distinguish beneficial financing from durable market foreclosure.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Investigate restrictive partnerships** [Near-term priority]. FTC and DOJ should use existing investigative and antitrust powers to examine exclusivity, cloud-spending commitments, and control rights, pursuing cases where evidence meets statutory tests. The FTC report identifies concrete mechanisms; enforcement must distinguish funding that enables entry from conduct that forecloses competitors.
2. **Review acquisitions of teams and assets** [Near-term priority]. Antitrust agencies should examine transactions that transfer a startup's core staff, intellectual property, or competitive capacity without a conventional acquisition. Congress should close reporting gaps if existing thresholds miss consequential deals; disclosures should be narrowly targeted so routine hiring does not trigger merger bureaucracy.
3. **Make cloud exit practical** [Structural reform]. Congress should authorize interoperability and reasonable data-egress rules for demonstrably powerful cloud gatekeepers, with standards developed through NIST and enforced by a designated regulator. Lower switching costs respond to

FTC concerns; mandates should protect security and legitimate migration costs rather than require identical proprietary architectures.

4. **Stop discriminatory model access** [Structural reform]. Congress should prohibit dominant input providers from unreasonably denying or degrading service to similarly situated downstream competitors, with expedited dispute resolution. Brookings identifies this supplier–competitor conflict; genuine safety violations, capacity limits, and intellectual-property protection require explicit, reviewable exceptions.
5. **Separate competitors' confidential data** [Near-term priority]. FTC settlements and federal procurement terms should require access controls and independent monitoring where providers also compete with customer firms. This addresses sensitive-information risks in the FTC report; rules should permit consensual service improvement while blocking covert use of rivals' commercial secrets.
6. **Buy from multiple suppliers** [Near-term priority]. Federal procurement agencies should require portable interfaces and credible exit plans for major AI contracts, and avoid unjustified exclusive commitments. Diversification creates openings for smaller firms and limits dependence; it adds integration costs and should not force agencies to use inadequately tested systems.
7. **Fund public-interest compute** [Near-term priority]. Congress should give NSF a stable appropriation for competitively allocated research computing, with independent panels and support for less-resourced universities. Industry-influence research supports this access intervention; allocation must avoid becoming a subsidy to providers selected through opaque partnerships.
8. **Publish concentration evidence** [Near-term priority]. Congress should fund FTC market studies tracking chip supply, cloud dependence, model access, and downstream competition with protected confidential submissions. Layer-specific evidence is more useful than a single chatbot market-share figure; publication should use aggregated measures and identify uncertainty about market definitions.
9. **Create interoperable data cooperatives** [Structural reform]. States and Congress should authorize voluntary, privacy-protective collective bargaining over clearly controlled datasets, including standardized licenses and member governance. Shared bargaining can improve participants' terms; antitrust safeguards, consent, and clear rights are essential because a cooperative cannot license data it does not lawfully control.
10. **Apply conditions to public subsidies** [Near-term priority]. Commerce and grant-making agencies should attach enforceable public-return provisions to major AI infrastructure subsidies, including clawbacks for unmet commitments and negotiated revenue sharing where lawful. Public investment should produce public benefits; demanding excessive returns may deter applicants or shift projects abroad.
11. **Tax economic rents more effectively** [Structural reform]. Congress should ask Treasury to design reforms addressing profit shifting and excess returns while preserving deductions for real investment, with distributional and revenue estimates. Ownership gains need not reach workers automatically; distinguishing rents from normal returns is difficult, so proposals require clear taxable bases rather than an undefined “AI wealth tax.”
12. **Support worker and cooperative ownership** [Near-term priority]. SBA and state development agencies should finance employee buyouts and cooperative AI-service businesses through loans with ordinary credit review and technical assistance. Broader ownership can distribute gains, but governance costs and failure risk mean these models should compete on viable business plans rather than receive permanent operating protection.
13. **Protect public-benefit governance** [Structural reform]. State attorneys general and legislatures should require documented conflict procedures and enforceable mission commitments when nonprofit-controlled AI assets are restructured. OpenAI's restructuring illustrates the stakes; stronger duties must specify who can enforce them and should not imply that nonprofit status itself proves safe conduct.
14. **Expose regulatory conflicts** [Near-term priority]. Congress should require public disclosure of advisory-panel financial interests, recusal rules, and publication of major regulatory meetings, with

protected treatment of security-sensitive material. Reporting on lobbying supports scrutiny of influence; public-interest groups and academics need comparable disclosure standards, not immunity from them.

15. **Test a public wealth fund** [Long-term proposal]. Congress should study and, if justified, pilot an independently governed fund holding diversified assets or negotiated public equity stakes, with benefits distributed under statutory rules. This could broaden participation in capital gains; market risk, politicized investments, and capture require strict fiduciary duties and no presumption of guaranteed AI windfalls.

E. STRONGEST COUNTERARGUMENTS

"Scale finances research and safety." Large integrated firms can fund expensive infrastructure and dependable services. Enforce evidence-backed restraints and improve access rather than breaking firms up on size alone; the durable relationship between concentration, innovation, and safety remains unsettled.

"New models will disrupt today's leaders." Efficiency improvements and entry are real, making permanent-monopoly claims premature. Competition at the chatbot layer does not eliminate cloud or chip dependence, so monitor each layer and target identifiable barriers; future switching and entry rates will determine how strong intervention should be.

"Redistribution discourages investment." Taxes and equity conditions can change incentives, and public funds can lose money. Favor transparent, predictable rules and measured returns for public subsidies; the appropriate tax design depends on evidence about rents, avoidance, and investment responses.

F. AI ACTION ADVOCACY BRIEF

Core message: AI's gains should reach the public, and a handful of suppliers should not decide who gets to compete.

Three strongest talking points

- Partnerships can fund new competitors and create dependence at the same time.
- Competition must be assessed across chips, cloud, models, and applications.
- Public research access and workable exit rights can support innovation as well as accountability.

Three priority policy proposals

- Evidence-led review of restrictive partnerships (1).
- Public-interest computing access (7).
- Protection against unreasonable denial of essential model access (4).

Questions for lawmakers

- Which partnership terms will antitrust agencies investigate, and what staffing will they have?
- Can publicly funded researchers access sufficient compute without depending on a single vendor?
- What enforceable public return accompanies your proposed infrastructure subsidies?

What we still don't know

- Which current advantages will persist as models become cheaper and easier to substitute.
- Whether supplier control raises prices or suppresses downstream innovation in particular markets.
- Which ownership and tax reforms share gains most effectively without excessive administrative costs.

EXISTENTIAL RISK AND LOSS OF HUMAN CONTROL

Central question: How can humanity maintain control over increasingly capable AI?

A. ISSUE OVERVIEW

- Loss of control means a system persistently pursues behavior its operators cannot reliably constrain, correct, or stop. An existential catastrophe is a much stronger claim: an outcome that destroys humanity or permanently undermines its future; neither follows simply because a chatbot makes a mistake.
- More capable agents can use software, obtain permissions, and coordinate actions across systems. That creates a political obligation to test serious hazards before giving them access to consequential infrastructure, even though the probability and timing of extreme outcomes remain disputed.
- Scientific research, medicine, and cybersecurity may benefit enormously from capable AI. The public bears risks that companies cannot fully contain within their own facilities, while premature restrictions can also concentrate power and slow beneficial research.
- Laboratory studies have documented deceptive behavior under deliberately constructed conditions. Independent evaluators measure growing ability to complete longer software tasks, while warning against treating benchmark extrapolations as certain futures. The 2026 International AI Safety Report reviews evidence and unresolved disagreement about loss-of-control pathways. California's 2025 frontier-AI transparency law provides an enacted reporting precedent, and 2026 reporting describes new company-disclosed incidents and proposed safeguards that still require independent scrutiny.
- The hardest question is whether present warning signs reliably predict dangerous behavior in deployment. Another is which safeguards can stop a highly capable system in practice without handing governments or incumbent companies sweeping, unaccountable control over ordinary AI development.

B. KEY FACTS AND EVIDENCE

- A 2024 Science paper by Bengio and colleagues argues for precaution against extreme risks. It is a consensus policy argument, not an empirical estimate of humanity's extinction probability. Science paper.
- Anthropic and Redwood's constructed alignment-faking experiment found harmful compliance with alignment-faking reasoning in 12% of cases in one monitored condition. The setup included a fictional training objective and hidden scratchpad; it does not estimate real-world prevalence. Alignment-faking study.
- METR's original software-task study estimated an approximately seven-month doubling time for its historical task-horizon metric. That trend is conditional on the task suite and models studied, not a deadline for general autonomy. Task-horizon research.
- METR's “50% time horizon” is the human-estimated task length at which an agent succeeds about half the time. It should not be described as reliable performance at every task of that duration. Metric definition.
- The 2026 International AI Safety Report treats general-purpose AI risks and mitigation as active research questions. Scientific synthesis supports preparation, while significant uncertainty remains about the pathways and likelihood of catastrophic loss of control. International report.
- California enacted SB 53 in 2025, requiring covered frontier developers to disclose safety frameworks, report certain critical incidents, and protect whistleblowers. It provides an oversight precedent, not a guarantee that dangerous behavior is prevented. Enacted California law.

C. RESEARCH AND SOURCE LIBRARY

1. Peer-reviewed academic research

Managing extreme AI risks amid rapid progress — Yoshua Bengio, Geoffrey Hinton, Andrew Yao and 22 coauthors; 2024; Science; DOI 10.1126/science.adn0117. A peer-reviewed policy perspective proposes testing, security, and adaptive governance for extreme risks. It articulates precautionary reasoning rather than measuring the probability of catastrophe.

2. Government and intergovernmental reports

International AI Safety Report 2026 — international expert panel chaired by Yoshua Bengio; February 2026. A government-supported scientific synthesis examines capabilities, misuse, malfunctions, and systemic risks. Use its distinctions and uncertainty to evaluate policy; it is not a binding international agreement.

SB 53 signing and policy summary — Governor of California; September 2025; official implementation context. Describes enacted disclosure, reporting, and whistleblower provisions. Useful as a legislative precedent; the governor's claims about economic benefits are advocacy, not independent evaluation.

3. AI industry research

Alignment faking in large language models — Anthropic and Redwood Research; December 2024; technical research initially released as a preprint. Demonstrates strategically different behavior under a contrived training setup. It motivates adversarial testing, with limited external validity and a company interest in how its safety work is perceived.

4. Independent think tanks and nonprofit organizations

Measuring AI Ability to Complete Long Software Tasks — Thomas Kwa, Ben West, Joel Becker and colleagues; METR; 2025. Develops human-calibrated autonomy benchmarks and conditional forecasts. Its policy contribution is a measurable evaluation approach; software benchmarks cannot alone determine acceptable social risk.

5. Credible journalism

Anthropic's Mythos model found vulnerabilities in classified US government systems, official says — AP; June 2026. Reports a testing claim attributed to an unnamed U.S. official. Finding vulnerabilities is different from exploiting them; the report is a reason for scrutiny, not independent verification of all capabilities.

OpenAI flags concerning new AI behavior and vows to track it more closely — AP; September 2026. Covers company-disclosed anomalous behavior and a monitoring framework. Highlights the importance of incident reporting while leaving underlying claims open to external review.

Anthropic CEO Dario Amodei says AI industry needs to give safety measures time to catch up — AP; September 2026. Reports an executive's slowdown proposal and competing incentives. Treat the forecast as the CEO's judgment, not established scientific timing.

Nvidia is touting a software tool to contain runaway AI. How would it work? — AP; September 2026. Explains a proposed agent-security platform and its context. Makes containment mechanisms concrete; an announced product is not proof that containment works.

6. Strongest opposing perspective

AI as Normal Technology — Arvind Narayanan and Sayash Kapoor; April 2025; research-informed essay. Argues that adoption, institutions, and engineering constraints make abrupt autonomous takeover narratives less persuasive than advocates suggest. Independent tests of sustained autonomy, adaptation outside benchmarks, and effective containment would help distinguish the competing accounts.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Require dangerous-capability evaluations** [Near-term priority]. Congress should give a designated federal safety authority power to require secure testing of covered frontier systems for autonomous cyber misuse, biological assistance, and evasion of control, using NIST-developed protocols. METR shows that capability can be measured, but tests need frequent revision and confidence intervals rather than a single pass–fail score.
2. **Demand a deployment safety case** [Structural reform]. Covered developers should submit an evidence-backed argument explaining threats, mitigations, residual risk, and authorized uses before consequential deployment, subject to independent review. The Science perspective supports proactive governance; reviewers must be technically competent and cannot treat a polished document as proof of safety.
3. **Set conditional deployment limits** [Structural reform]. Congress should authorize restrictions when validated tests reveal specified dangerous capabilities and safeguards are inadequate, with narrow scope, public reasons, and appeal rights. This targets credible hazards rather than banning all

advanced AI; measurement error and uncertain benefit–risk comparisons make periodic review essential.

4. **Retest material changes** [Near-term priority]. The same statutory regime should require renewed assessment after substantial fine-tuning, new tools, added autonomy, or deployment in a new critical setting. Alignment-faking research motivates testing behavior across conditions; proportional rules should exempt changes shown not to alter the relevant risk profile.
5. **Secure high-risk model assets** [Near-term priority]. Congress should authorize audited security standards for weights, privileged access, and critical infrastructure at developers of demonstrably dangerous systems, drawing on established cybersecurity practice. Preventing theft and unauthorized modification limits proliferation; audit summaries should preserve security-sensitive information and avoid imposing frontier-grade requirements on ordinary research models.
6. **Limit agents' privileges** [Near-term priority]. Federal contracts and sector regulators should require least-privilege access, explicit spending limits, network restrictions, and human authorization for irreversible actions by high-impact agents. These barriers reduce the consequences of errors or deception; access controls need realistic adversarial tests and can reduce useful automation.
7. **Keep auditable action records** [Near-term priority]. Covered deployers should retain tamper-evident logs of consequential agent actions and permission changes, accessible to authorized investigators under privacy protections. Observable actions are more dependable evidence than a model's self-explanation; retention periods and access controls should prevent surveillance of users.
8. **Report critical incidents and near misses** [Near-term priority]. Congress should establish a protected reporting channel with deadlines proportionate to severity, modeled in part on SB 53, and publish de-identified incident patterns. Shared evidence can reveal recurring control failures; rules should distinguish genuine emergencies from harmless anomalies and prevent disclosure of exploitable details.
9. **Protect safety whistleblowers** [Near-term priority]. Federal law should prohibit retaliation for good-faith reports of serious AI safety misconduct to authorized regulators, provide confidential channels, and preserve trade-secret remedies for unrelated disclosures. SB 53 offers a precedent; investigators need access to evidence without encouraging public release of dangerous assets.
10. **Enable independent red teaming** [Structural reform]. Congress should create a narrow safe harbor for authorized safety testing and require meaningful evaluator access under secure agreements for covered systems. Company studies alone cannot settle deployment risk; safeguards must protect personal data and avoid immunizing extortion, unauthorized intrusion, or reckless publication.
11. **Build public evaluation capacity** [Near-term priority]. Congress should fund technical staff, secure compute, benchmark maintenance, and independent laboratory partnerships at the federal safety authority. Independent scrutiny makes safety obligations enforceable; procurement conflicts and dependence on developers' staff require public conflict rules and evaluator rotation.
12. **Fund control and interpretability research** [Near-term priority]. NSF and other research agencies should support competitive projects testing monitoring, corrigibility, anomaly detection, and containment, with reproducible results wherever safe. Existing experiments show unsolved questions rather than validated universal fixes; funding should include skeptical replication and avoid promising a guaranteed technical solution.
13. **Create reviewable emergency suspension powers** [Structural reform]. Congress should authorize time-limited suspension of a particular deployment when credible evidence indicates imminent severe harm, with independent technical review and rapid judicial challenge. Emergency authority can preserve response time; broad executive discretion could suppress research or dissent, so triggers and renewal conditions must be explicit.
14. **Agree on cross-border testing and incident exchange** [Near-term priority]. State Department and technical agencies should negotiate shared evaluation protocols, secure notification channels, and joint exercises with willing countries, including competitors where feasible. The international report demonstrates scientific cooperation; verification, espionage concerns, and unequal state capacity limit what voluntary agreements can achieve.
15. **Consider licensing only at demonstrated extreme-risk thresholds** [Long-term proposal]. Congress should consider a license for development or operation of systems crossing independently

validated dangerous-capability thresholds, contingent on security, evaluation access, and credible control measures. Compute may help identify candidates but should not be the sole trigger; licensing risks capture, offshore development, and barriers to beneficial entrants. See Tab 4 for wider regulatory design.

E. STRONGEST COUNTERARGUMENTS

"The catastrophe case extrapolates too far." Constructed experiments and software benchmarks do not establish a real-world takeover pathway or a calibrated probability. Favor observable tests, access controls, and incident reporting while keeping severe-risk restrictions conditional; the evidence linking current behaviors to existential outcomes remains incomplete.

"Safety rules will entrench incumbents." Evaluations and security can be expensive, and companies may lobby for rules competitors cannot satisfy. Use public testing support, transparent thresholds, and exemptions tied to capability and use; some expensive safeguards may still be justified for exceptionally dangerous systems.

"Global enforcement is impossible." Models, expertise, and compute cross borders, while governments have conflicting interests. Shared protocols and security can reduce risk without perfect agreement, but neither a domestic law nor voluntary diplomacy can guarantee universal compliance.

F. AI ACTION ADVOCACY BRIEF

Core message: Powerful AI should have to demonstrate workable safeguards before it receives dangerous levels of access and autonomy.

Three strongest talking points

- Warning signs justify testing; they do not establish a numerical probability of extinction.
- Independent evaluation must examine behavior across conditions, not just benchmark scores.
- Reporting failures and limiting permissions can help now while longer-term risks are investigated.

Three priority policy proposals

- Mandatory dangerous-capability evaluations (1).
- Critical-incident and near-miss reporting (8).
- Least-privilege controls for consequential agents (6).

Questions for lawmakers

- Which agency can compel independent testing, and will you fund its technical staff?
- What evidence would justify restricting a specific deployment, and what appeal protects against abuse?
- Who receives serious incident reports, and what protects an employee who reports them?

What we still don't know

- How reliably present experiments predict behavior outside the laboratory.
- Which containment and monitoring tools remain effective as systems become more capable.
- The likelihood, timing, and concrete pathways of catastrophic loss of control.

AI REGULATION VERSUS INNOVATION

Central question: How much freedom should AI developers have?

A. ISSUE OVERVIEW

- AI governance is a choice about obligations, enforcement, and responsibility across different products and uses. A writing assistant, a hiring filter, and a highly autonomous frontier system do not create the same risks, so an effective framework should explain which rules apply to each.
- Good regulation can protect rights and make markets more trustworthy. Poorly designed rules can raise costs without reducing harm, protect incumbents, or transfer too much authority to officials and industry-selected auditors.
- Users, workers, and businesses benefit when unsafe conduct has consequences and compliant products have clear paths to market. Startups can benefit from predictable standards but struggle with fixed compliance costs; large firms can absorb those costs and may help shape rules around their own practices.
- The United States combines existing sector laws with state AI measures and continuing arguments over federal authority. NIST's risk framework remains voluntary guidance rather than a universal license. California enacted frontier transparency requirements in 2025, while a proposed congressional moratorium on state AI rules was removed that year. The EU uses a risk-based statute, but its 2026 amendments changed important high-risk implementation dates, showing why organizers must distinguish enacted duties, future deadlines, and voluntary codes.
- The unresolved questions are which obligations demonstrably improve outcomes and who should enforce them. Another is how to provide consistent national rules while preserving enforceable protections, public participation, and room for useful experiments by smaller developers.

B. KEY FACTS AND EVIDENCE

- NIST's AI Risk Management Framework 1.0 is voluntary and organized around governing, mapping, measuring, and managing risk. Claiming that an organization “uses NIST” does not establish legal compliance or independent certification. NIST framework.
- The EU AI Act differentiates prohibited practices, high-risk uses, transparency duties, and general-purpose models. The categories concern different actors and obligations; there is no single blanket approval process for all AI. EU framework.
- The European Commission reports that the AI Omnibus entered into force in July 2026, extending Annex III high-risk rules to December 2027 and relevant product-embedded rules to August 2028. Earlier summaries using only the original timetable are outdated. Official amendments summary.
- California's SB 53, enacted in 2025, uses frontier transparency, incident reporting, and whistleblower provisions. Those duties differ from a license to develop every model. California precedent.
- The U.S. Senate voted 99–1 to remove a proposed state-AI regulatory restriction from the 2025 reconciliation bill. The rejected provision must not be described as enacted federal preemption. Senate reporting.
- The White House's March 2026 National AI Legislative Framework describes proposals for Congress. A framework announcement is not itself a statute and does not independently repeal state protections. Federal legislative proposal.

C. RESEARCH AND SOURCE LIBRARY

1. Peer-reviewed academic research

Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing — Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White and colleagues; 2020; ACM FAccT proceedings, not a journal. Proposes lifecycle auditing rather than a one-time technical check. This peer-reviewed framework supports documentation and responsibility, while not proving that audits reduce all harms.

2. Government and intergovernmental reports

Artificial Intelligence Risk Management Framework (AI RMF 1.0) — Elham Tabassi; NIST; 2023. A voluntary, rights-preserving risk-management framework. It supplies operational vocabulary that lawmakers can adapt into specific duties, rather than a ready-made statutory enforcement regime.

AI Omnibus enters into force — European Commission; July 2026; official regulatory summary. Explains revised deadlines and administrative simplification. Use it with the Commission's AI Act guidance to distinguish current obligations from delayed high-risk requirements.

3. AI industry research

Model Cards for Model Reporting — Margaret Mitchell, Simone Wu, Andrew Zaldivar and colleagues; Google-associated research; 2019. Proposes standardized documentation of intended uses, performance, and limitations, illustrated with model examples. It is technical governance research, not evidence that documentation alone prevents harm; industry authors have implementation and reputational interests.

4. Independent think tanks and nonprofit organizations

The three challenges of AI regulation — Tom Wheeler; Brookings; 2023. Argues for risk-matched rules, adaptable standards, and institutional capacity. Its practical contribution is regulatory design that can evolve without leaving all enforcement to voluntary commitments.

5. Credible journalism

Senate strikes AI regulatory ban from GOP bill after uproar from the states — AP; July 2025. Reports the defeat of a moratorium proposal and its political coalition. Useful for assessing federalism and feasibility.

California Gov. Gavin Newsom signs landmark bill creating AI safety measures — AP; September 2025. Examines SB 53's passage and competing claims about innovation and safety. It documents enactment rather than a completed evaluation of effects.

Trump tried to block state AI regulations, but some states are forging ahead — AP; June 2026. Reports continued state activity and the dispute over federal intervention. Shows how authority and policy coverage remain contested.

EU to crack down on AI deepfakes, illicit imagery and hacking with new team in Brussels — AP; July 2026. Covers new enforcement capacity and implementation. Pair it with official timelines, since different parts of the Act apply at different dates.

6. Strongest opposing perspective

The Case for Smarter AI Regulation, With Matt Perault — ITIF podcast; July 2025; featuring Andreessen Horowitz's AI policy lead. Makes the case that broad development and disclosure mandates burden startups and that regulation should focus on harmful uses. This is an interested industry perspective; comparative evidence on compliance costs, entry, incidents, and enforcement outcomes would test it.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Regulate high-impact uses proportionately** [Structural reform]. Congress should define consequential uses in employment, credit, housing, health, and public services, then assign testing, recordkeeping, and appeal duties to the responsible actors. The EU's differentiated structure offers a precedent; a risk label must reflect actual impact and avoid forcing every small application through the same process.
2. **Enforce existing laws with technical staff** [Near-term priority]. Congress should fund AI expertise at civil-rights, consumer-protection, and sector regulators, using their existing statutory powers while clearly identifying gaps requiring legislation. NIST can inform technical practice; agencies must not assume guidance creates authority Congress has not granted.
3. **Require concise system documentation** [Near-term priority]. Congress or authorized regulators should require model-card-style disclosure of intended uses, limitations, test conditions, and material changes for covered high-impact systems. Model-card research supplies a practical format;

confidential annexes and standardized templates can reduce trade-secret exposure and small-firm costs.

4. **Assign developer and deployer duties** [Structural reform]. Legislation should make developers responsible for reasonably foreseeable design failures and deployers responsible for inappropriate use, local validation, and oversight, without contractual evasion of statutory rights. Lifecycle auditing supports this division; liability must reflect control, knowledge, and causation rather than automatically blaming every supplier for every misuse.
5. **Audit where independent access is needed** [Near-term priority]. Authorized regulators should require qualified, conflict-managed audits for high-risk systems, including deployment context, sampling, and access to relevant evidence. Raji and colleagues show why audits should cover the lifecycle; auditor accreditation, rotation, and quality checks are needed to prevent a box-checking market.
6. **Create targeted regulatory sandboxes** [Near-term priority]. Sector agencies should offer supervised, time-limited pilots with defined participants, safeguards, outcome reporting, and exit criteria. Brookings' adaptable approach supports experimentation; participation must not waive civil rights or turn pilot status into an indefinite exemption from liability.
7. **Help smaller firms comply** [Near-term priority]. Congress should finance public testing facilities, reusable documentation templates, and technical assistance for eligible small developers. Fixed-cost concerns raised by Perault deserve attention; assistance should be independent of major vendors and should not lower substantive protection for people exposed to high-risk products.
8. **Publish incident-based enforcement priorities** [Near-term priority]. Regulators should publish annual priorities based on verified complaints, incidents, and exposure, while allowing confidential reporting of serious vulnerabilities. This makes risk management responsive to evidence; low complaint volumes may reflect barriers to reporting rather than absence of harm.
9. **Give affected people enforceable remedies** [Structural reform]. Congress should provide notice, meaningful contestation, and suitable remedies for consequential automated decisions, alongside regulator enforcement. Audits alone cannot compensate individuals or correct a particular decision; procedure needs deadlines and accessible explanations, with details tailored to each sector. See Tabs 12 and 15.
10. **Keep federalism protections explicit** [Structural reform]. Any federal AI statute should state precisely which requirements it preempts and preserve stronger state protections unless Congress supplies enforceable equivalent safeguards. The 2025 moratorium debate illustrates the stakes; national consistency can reduce costs, but preemption without effective enforcement can remove protection altogether.
11. **Review rules using measured outcomes** [Near-term priority]. Congress should require periodic evaluation of incident rates, compliance costs, entry, and access, with public consultation and statutory authority for justified revisions. The EU's revised timetable shows rules can change; review should not automatically expire protections before evidence can be collected.
12. **Guard against regulatory capture** [Near-term priority]. Regulators should disclose relevant financial interests, publish meeting summaries, and ensure advisory bodies include independent scientists, smaller firms, affected communities, and civil-rights expertise. Technical standards can otherwise reflect incumbents' practices; representation is useful only if decision-making and evidence remain transparent.
13. **Align interoperable standards internationally** [Near-term priority]. Commerce, NIST, and State should negotiate common test and documentation formats while preserving each jurisdiction's substantive rights. Shared standards can reduce duplication highlighted by innovation critics; mutual recognition should depend on demonstrably comparable oversight rather than accepting the weakest jurisdiction's certificate.
14. **Use licensing only for exceptional hazards** [Long-term proposal]. Congress should consider licensing systems with independently validated catastrophic-risk capabilities, defining thresholds and appeal rights rather than licensing all developers. This precaution is stronger than routine use-based

regulation and requires evidence of otherwise unmanageable risk; fees and discretionary approvals can entrench incumbents. See Tab 3.

15. **Test governance capacity before creating a new regulator** [Structural reform]. Congress should commission a time-limited assessment of agency gaps and then establish a specialized body only if existing regulators cannot address cross-sector risks, with clear jurisdiction and funding. Institutional capacity is central to Brookings' argument; overlapping mandates and a captured central regulator could make oversight worse.

E. STRONGEST COUNTERARGUMENTS

"Compliance will favor large firms." Fixed costs and uncertain obligations can deter entry. Public testing, clear templates, and risk-based duties address this credible objection, but some costs remain unavoidable for dangerous uses and should be measured rather than assumed harmless.

"Regulators cannot keep up." Technology changes faster than legislation, and inexperienced officials may approve ineffective paperwork. Stable legal duties combined with revisable technical standards and funded expertise are more credible than prescribing one model architecture; the evidence on which governance tools work remains limited.

"A national rule should replace the patchwork." Inconsistent requirements can consume resources and complicate deployment. Targeted preemption and common standards can help, while retaining remedies and stronger protections; the best allocation of authority depends on the actual federal statute and enforcement capacity.

F. AI ACTION ADVOCACY BRIEF

Core message: Set clear rules for serious risks, enforce them, and make compliance workable for smaller firms.

Three strongest talking points

- Voluntary risk-management guidance is useful but does not guarantee accountability.
- Different uses require different duties and different regulators.
- Safety and competition can improve together when rules avoid unnecessary fixed costs.

Three priority policy proposals

- Proportionate duties for high-impact uses (1).
- Funded enforcement of existing rights (2).
- Shared testing and compliance support for small firms (7).

Questions for lawmakers

- Who enforces each proposed duty, and what evidence must a covered company provide?
- What will your rules cost a small developer with a genuinely high-risk product?
- If you preempt state law, which enforceable federal protections replace it?

What we still don't know

- How much specific audits and disclosures reduce incidents in practice.
- Which requirements create material barriers to entry rather than manageable costs.
- Whether current institutions can oversee cross-sector frontier risks effectively.

DEMOCRACY, ELECTIONS, AND MISINFORMATION

Central question: Can democratic institutions function when convincing political content can be manufactured at scale?

A. ISSUE OVERVIEW

- Generative AI lowers the cost of producing political text, fake accounts, and realistic audio or video. It can also help citizens understand policy, translate official information, and challenge false claims; the political consequences depend on distribution, incentives, and audience response.
- Democracy requires reliable voting information and the freedom to argue, criticize, and organize. Deceptive impersonation and fabricated procedures can undermine those conditions, while sweeping bans on “misinformation” can empower officials to suppress legitimate opponents.
- Campaigns and small civic groups benefit from cheaper communication tools. Voters, journalists, election workers, and targeted communities may face manipulation, harassment, and greater difficulty proving that authentic evidence is genuine.
- U.S. states have enacted different disclosure and deepfake measures, but 2026 reporting identifies continuing gaps in political-ad labeling. Federal robocall rules cover AI-generated voices under existing artificial-voice restrictions, while an October 2026 waiver request remains a proposal in the reporting reviewed. Moldova and Brazil provide recent examples of foreign-influence concerns and disputed political imagery. Research demonstrates persuasion in controlled conversations, but analyses of past elections caution that the presence of synthetic media does not by itself establish changed election results.
- The most important unknown is how laboratory persuasion translates into repeated exposure, real campaigns, and actual voting. Another is which safeguards protect voters while preserving satire, anonymity, criticism, and a politically neutral process for correcting dangerous falsehoods.

B. KEY FACTS AND EVIDENCE

- A preregistered study of 900 participants found personalized GPT-4 could outperform human opponents in short debates. It measured post-debate agreement, not votes or election outcomes. Persuasion experiment.
- A separate experiment involving 2,190 conspiracy believers found AI dialogue reduced belief by about 20%, with persistence measured afterward. AI can support accurate information as well as manipulation; generalization beyond this intervention remains uncertain. Conspiracy-belief study.
- The FCC's 2024 ruling classifies AI-generated voices as “artificial” under the TCPA. Consent requirements, exemptions, and call context matter; it is inaccurate to say every AI voice call is illegal. FCC declaratory ruling.
- OpenAI's October 2025 report says it had disrupted more than 40 policy-violating networks since public threat reporting began. That company count includes several types of abuse and does not measure election effects or all campaigns. Threat-intelligence report.
- AP reported in September 2026 that the Wesleyan Media Project had identified at least 164 political ads generated or enhanced with AI in the cycle. “AI-enhanced” is broader than deceptive, and this reporting does not establish how many votes were changed. Ad investigation.
- An analysis of 78 documented 2024 election deepfakes found reasons to question claims of an AI-driven misinformation apocalypse. Its selected, publicly observed cases cannot rule out hidden activity or future effects. Opposing case analysis.

C. RESEARCH AND SOURCE LIBRARY

1. Peer-reviewed academic research

On the conversational persuasiveness of GPT-4 — Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, Robert West; 2025; Nature Human Behaviour. Controlled debates test personalization and

persuasive effects. Supports scrutiny of targeted political agents, with no direct estimate of effects on turnout or elections.

Durably reducing conspiracy beliefs through dialogues with AI — Thomas H. Costello, Gordon Pennycook, David G. Rand; 2024; Science; DOI 10.1126/science.adq1814. Evidence-based dialogue reduced reported conspiracy belief. Demonstrates a beneficial intervention while leaving open how it performs at public scale and across populations.

2. Government and intergovernmental reports

FCC 24-17: Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts — FCC; February 2024; declaratory ruling. Applies existing artificial-voice restrictions to AI-generated speech. It provides an enforcement mechanism, not a general rule for all online political content.

3. AI industry research

Disrupting malicious uses of AI: October 2025 — OpenAI; October 2025; threat-intelligence research. Documents identified abuse networks, including covert influence activity. Findings depend on the company's visibility and attribution methods, and claims of successful disruption are not independently verified outcomes for democracy.

4. Independent think tanks and nonprofit organizations

An Agenda to Strengthen U.S. Democracy in the Age of AI — Mekela Panditharatne, Lawrence Norden, Joanna Zdanys, Daniel I. Weiner, Yasmin Abusaif; Brennan Center; February 2025. Offers election, government-accountability, and platform-policy proposals. Useful for translating concerns into disclosure, preparedness, and enforceable public safeguards.

5. Credible journalism

AI ads are popular in the 2026 midterms, but state laws haven't stopped a lack of disclosure — AP; September 2026. Investigates political ads and practical disclosure gaps. Shows why a rule's existence is different from enforcement.

Moldova's election faces AI-driven disinformation from Russia — AP; September 2025. Reports monitoring groups' findings about coordinated influence campaigns. Distinguish attributed activity from demonstrated electoral impact.

Brazil's electoral court sets new standards for AI deepfakes after case involving Sen. Bolsonaro — AP; September 2026. Examines a divided court's treatment of synthetic campaign imagery. Useful for questions about deception, artistic portrayal, and political expression.

FCC considers allowing more political robocalls before midterms, including those created with AI — AP; October 2026. Reports a waiver request and objections concerning unwanted calls and privacy. The request is not described here as an approved rule change.

6. Strongest opposing perspective

We Looked at 78 Election Deepfakes. Political Misinformation Is Not an AI Problem. — Sayash Kapoor and Arvind Narayanan; Knight First Amendment Institute; December 2024. Argues that political incentives and information distribution matter more than the production technology in observed cases. Representative exposure data and credible causal studies of behavior would test the disagreement; a case collection cannot settle all future risks.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Label materially deceptive campaign media** [Near-term priority]. Congress should amend federal campaign-disclaimer law to require clear disclosure when paid campaign media materially fabricates an identifiable person's words or actions, with FEC enforcement and state counterparts. Brennan's proposals support this narrow approach; ordinary editing, obvious satire, and accessibility tools need precise exceptions.

2. **Enforce voice-cloning call restrictions** [Near-term priority]. FCC and state attorneys general should enforce applicable TCPA consent and identification rules against deceptive AI robocalls, with prompt complaint handling. The 2024 ruling supplies existing authority; enforcement must respect lawful exceptions and distinguish a proposed waiver from an adopted change.
3. **Prohibit fraudulent voting instructions** [Structural reform]. Congress should create a narrowly defined federal remedy for knowingly deceptive statements about voting eligibility, time, place, or method made with intent to prevent lawful voting. This addresses a specific democratic harm rather than general political falsity; intent, materiality, and rapid judicial review are essential speech safeguards.
4. **Provide rapid, narrow court remedies** [Structural reform]. Election laws should permit expedited proceedings against actionable impersonation or voting fraud, with evidence requirements, notice, and orders limited to the harmful material. Disclosure gaps justify a usable enforcement path; pre-election injunctions can chill speech, so courts must protect criticism and provide prompt appeal.
5. **Make political-ad archives searchable** [Near-term priority]. Congress should require major ad platforms to retain sponsor, spending range, dates, broad targeting criteria, and disclosed synthetic-media use in a public research archive. The 2026 ad investigation supports better visibility; records should protect individual users and avoid publishing sensitive personal profiles.
6. **Limit sensitive-data political targeting** [Structural reform]. Federal privacy legislation should prohibit political advertisers from using unlawfully obtained data and restrict targeting based on highly sensitive personal attributes without meaningful consent. The persuasion study makes personalization relevant; restrictions must define data uses rather than suppress a campaign's viewpoint or lawful broad geographic outreach.
7. **Identify automated campaign agents** [Near-term priority]. Congress should require interactive campaign agents to disclose automation and the sponsoring political organization at the start of a conversation. Persuasion research supports knowing who is speaking; a duty should cover organized campaign communication, preserve individual anonymity, and avoid forcing private citizens to disclose every writing tool.
8. **Authenticate official election communications** [Near-term priority]. Congress should fund election offices to publish signed, accessible announcements on verified domains, maintain callback procedures, and offer multilingual voting information. Brennan's preparedness approach supports this practical defense; authentication proves origin, not truth, and must work for people with limited internet access.
9. **Preserve useful provenance metadata** [Near-term priority]. Public procurement and voluntary platform standards should support interoperable content credentials and warn when signed provenance has been removed. This can help verify origin, but stripped metadata and unsigned authentic media are common; government should not make provenance a universal condition for lawful publication.
10. **Disrupt coordinated deceptive identities** [Near-term priority]. Platforms should publish and enforce rules against coordinated fake personas and covert sponsorship, with appeal and transparency reports, while government investigates illegal conduct under existing authority. OpenAI's reporting documents abuse patterns; oversight must target deceptive behavior rather than ideological agreement or anonymous dissent.
11. **Run election-response exercises** [Near-term priority]. EAC grants should fund scenario-based drills for impersonation, false closure announcements, account compromise, and voice fraud, including verification and correction procedures. Preparedness recommendations supply a practical basis; exercises should involve local communities and measure response times rather than create partisan threat lists.
12. **Give researchers protected platform access** [Structural reform]. Congress should authorize privacy-preserving access to political-ad and influence-operation data for qualified independent researchers, with security controls and publication protections. Conflicting assessments of electoral impact make independent evidence valuable; safeguards must prevent re-identification and covert political surveillance.

13. **Support verification capacity in local journalism** [Near-term priority]. Congress and nonpartisan funders should offer competitively awarded grants for forensic verification, secure evidence storage, and multilingual reporting, insulated from editorial direction. The documented reporting cases show the need for capable intermediaries; public funding must not reward political loyalty or certify official truth.
14. **Teach verification habits through public programs** [Near-term priority]. Education agencies and libraries should test instruction on checking primary sources, tracing sponsors, and confirming voting information through official channels. Costello's study suggests corrective engagement can help, but it does not prove every literacy course works; evaluate transfer to real behavior and avoid claiming visual deepfake detection is dependable.
15. **Coordinate cross-border incident disclosure** [Near-term priority]. State Department and election-security agencies should share evidence of unlawful foreign interference with democratic partners through secure, accountable channels and publish attribution confidence. Moldova reporting illustrates the concern; uncertain attribution, diplomatic escalation, and citizens' privacy require independent review and limits on data exchange.

E. STRONGEST COUNTERARGUMENTS

"AI has not been shown to decide elections." The strongest skeptical analyses correctly distinguish content creation from exposure and behavioral effect. Target proven fraud and improve evidence collection rather than claim a stolen election without causal evidence; future targeted agents may differ from 2024 memes.

"Deepfake rules will censor satire and opponents." Political speech deserves strong protection, and ambiguous bans invite abuse. Narrow paid-ad disclosure and intent-based voting-fraud rules are more defensible than a government power to remove all false claims; constitutional boundaries still require case-specific analysis.

"Labels and watermarks will not stop malicious actors." Bad actors can evade them, and audiences may ignore warnings. Combine provenance with sponsor transparency, trusted channels, and enforcement; the net effect on belief and behavior remains an empirical question.

F. AI ACTION ADVOCACY BRIEF

Core message: Voters should know who is speaking and be able to trust information about how to vote.

Three strongest talking points

- Controlled research demonstrates persuasion, but does not prove an election was changed.
- Existing call-fraud rules apply to relevant uses of AI-generated voices.
- Clear disclosure and reliable official channels can protect voters without policing political opinions.

Three priority policy proposals

- Narrow disclosure for materially deceptive paid campaign media (1).
- Authenticated, accessible election communications (8).
- Public political-ad archives (5).

Questions for lawmakers

- How will your deepfake proposal protect satire, criticism, and anonymous speech?
- Who can correct fraudulent voting instructions quickly, and with what judicial safeguards?
- Will researchers and voters be able to see who paid for targeted political ads?

What we still don't know

- How personalized AI changes voting behavior outside controlled experiments.
- Whether labels and provenance tools improve decisions in actual campaigns.
- The scale of covert activity that providers and public researchers cannot observe.

SURVEILLANCE AND PRIVACY

Central question: Will AI make mass surveillance significantly easier?

A. ISSUE OVERVIEW

- AI can connect faces, locations, purchases, messages, and behavioral patterns across large datasets. It changes the cost and reach of surveillance, even when the underlying cameras, commercial records, and collection powers already exist.
- Privacy protects more than secrets: it gives people room to organize, seek care, express unpopular views, and make mistakes without permanent profiling. Surveillance becomes politically dangerous when people cannot discover who tracks them, contest an identification, or limit later uses of their data.
- Carefully bounded tools can help find a missing person, secure an account, or investigate a serious crime. Broad collection can expose bystanders, activists, immigrants, and vulnerable communities to tracking and wrongful inference, while commercial users may trade away information without understanding the consequences.
- U.S. reporting describes AI-assisted immigration enforcement, proposed transit surveillance, and expanding biometric systems. Canada has tested face-recognition body cameras against a watch list, illustrating how identification can move from retrospective investigation toward continuous monitoring. FTC research documents extensive commercial collection, while technical studies show that models can retain recoverable training data. European privacy and AI rules offer different legal approaches, but improved accuracy alone does not answer whether a particular use of surveillance is justified.
- The unresolved questions concern both performance and power: how often systems misidentify people in actual deployment, and which surveillance uses a free society should permit even if technically accurate. Effective safeguards must address collection, access, purpose, retention, and remedies together.

B. KEY FACTS AND EVIDENCE

- NIST's 2019 face-recognition study found substantial demographic differences in false-positive rates across many tested algorithms, often by factors of 10 to 100. These are historical benchmark results, not a universal error rate for today's systems. NIST demographic study.
- NIST also found large variation across algorithms, with better-performing systems generally showing smaller demographic effects. An error measure must specify the model, task, threshold, and image conditions. NIST findings.
- Peer-reviewed researchers extracted hundreds of memorized text sequences from GPT-2, including personal information in training material. This demonstrates a privacy failure mode, not that every model exposes every training record. Training-data extraction study.
- FTC's 2024 social-media and streaming investigation documented extensive collection and limited user control among studied services. It describes specific company practices rather than proving every AI application uses the same data. FTC report.
- AP reported a Canadian body-camera trial using a watch list of roughly 7,000 people. Scale and operating context matter: live identification can affect many people beyond the intended suspects. Body-camera investigation.
- Differential privacy can mathematically limit information about an individual's inclusion in a training dataset when correctly implemented with a defined privacy budget. It does not legalize collection or protect against every downstream disclosure. DeepMind privacy research.

C. RESEARCH AND SOURCE LIBRARY

1. Peer-reviewed academic research

Extracting Training Data from Large Language Models — Nicholas Carlini, Florian Tramèr, Eric Wallace and colleagues; 2021; 30th USENIX Security Symposium proceedings, not a journal. Demonstrates recovery of memorized training examples from GPT-2. Supports data minimization and privacy testing while leaving model-specific prevalence uncertain.

2. Government and intergovernmental reports

Face Recognition Vendor Test Part 3: Demographic Effects — Patrick Grother, Mei Ngan, Kayee Hanaoka; NISTIR 8280; 2019. Evaluates demographic performance differences in identification and verification. Provides a testing precedent; current procurement needs current model and deployment evidence.

A Look Behind the Screens: Examining the Data Practices of Social Media and Video Streaming Services — FTC; September 2024. Investigates collection, monetization, automated systems, and user control. Supports purpose limits and deletion rights, with conclusions bounded by the companies studied.

3. AI industry research

Unlocking High-Accuracy Differentially Private Image Classification through Scale — Google DeepMind research team; 2022. Technical work examines improved privacy-preserving training for image models. Useful for specifying measurable privacy safeguards; it does not independently certify the company's commercial data practices.

4. Independent think tanks and nonprofit organizations

New Orleans Police Continuing to Use Live Face Recognition Despite City Law — ACLU; April 2026; advocacy investigation. Uses obtained records to argue that third-party access can evade restrictions in practice. Its concrete policy contribution is closing indirect-use loopholes and improving disclosure; disputed legal interpretations remain claims, not court findings.

5. Credible journalism

A city's push for facial recognition on public buses ignites debate over security and privacy — AP; June 2026. Examines Kansas City's proposed transit use and competing safety and privacy arguments. Shows why ordinary public movement deserves its own safeguards.

AI-powered police body cameras, once taboo, get tested on Canadian city's 'watch list' of faces — AP; December 2025. Investigates an Edmonton trial and expert objections. Distinguishes a test from evidence that live scanning is safe or necessary.

DHS ramps up surveillance in immigration raids, sweeping in citizens — AP; January 2026. Reports use of facial identification, license-plate records, and commercial location data. Makes the risks of combined datasets and bystander exposure concrete.

Detailed findings from AP investigation into how US tech firms enabled China's digital police state — AP; 2025. Traces technology supply relationships and surveillance applications. Illustrates international human-rights risks; individual companies' roles should be assessed from the reported evidence, not guilt by association.

6. Strongest opposing perspective

Testimony Before the House Oversight Committee on Commercial Use of Facial Recognition Technology — Daniel Castro; ITIF; January 2020. Argues that bans sacrifice beneficial identification, authentication, and security uses, favoring performance and use safeguards instead. Independent comparisons with less intrusive alternatives, operational error data, and evidence of enforceable limits would test whether a particular ban is warranted.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Adopt a federal data-minimization duty** [Structural reform]. Congress should require covered businesses to collect and retain only personal data reasonably necessary for specified services, enforced by FTC and state attorneys general. The FTC investigation supports moving beyond notice alone; exceptions need precision so essential fraud detection and research remain possible without becoming unlimited loopholes.
2. **Require warrants for sensitive-data purchases** [Structural reform]. Congress should require judicial authorization for government acquisition of defined sensitive location, communications, and biometric data when comparable direct collection would require strong legal process. AP's immigration reporting shows why purchased data matters; emergency exceptions should be narrow, logged, and reviewed after the fact.

3. **Restrict continuous public face tracking** [Structural reform]. Congress and states should prohibit indiscriminate live remote biometric tracking in ordinary public spaces, allowing only narrowly defined, time-limited exceptions subject to independent authorization. Transit and body-camera cases illustrate reach; exemptions must preserve consensual device authentication and avoid a blanket ban on every face-related technology.
4. **Demand corroboration before enforcement action** [Near-term priority]. Police agencies should prohibit arrest, detention, or adverse immigration action based solely on a face-recognition candidate match, with enforceable policies and disclosure to defendants. NIST's false-positive evidence supports treating matches as leads; additional evidence must be independently gathered rather than merely repeat the same flawed identification.
5. **Test systems under actual deployment conditions** [Near-term priority]. Agencies should require independent evaluation of the specific algorithm, image quality, threshold, watch-list size, and affected groups before procurement and after updates. NIST supplies a methodological basis; benchmarks cannot replace local testing, and small agencies need shared evaluation support.
6. **Close third-party surveillance loopholes** [Near-term priority]. Legislatures should make agency restrictions cover contractor and informal third-party searches undertaken for government benefit, with records of requests and penalties for evasion. The ACLU's New Orleans investigation motivates this proposal; rules should distinguish prohibited proxy surveillance from lawful receipt of a genuine independent witness tip.
7. **Set retention and deletion deadlines** [Near-term priority]. Agencies and covered businesses should adopt purpose-specific deletion schedules for bystander biometric records, raw surveillance feeds, and AI chat logs, subject to narrowly scoped evidence holds. FTC findings support stronger control over stored data; deletion must preserve due-process evidence and should include backups and contractors.
8. **Give people access and correction rights** [Structural reform]. Congress should provide accessible rights to learn which consequential profiles are held about them, correct inaccuracies, and request deletion where appropriate. Combined-data surveillance makes errors consequential; verified requests and limited investigation exemptions are needed without creating an impossible identity-verification burden.
9. **Require separate consent for model training** [Structural reform]. Privacy legislation should require meaningful, specific consent for secondary training on sensitive private communications, with easy withdrawal for future use and clear limits on what deletion can technically achieve. Training-data extraction research supports caution; laws should not promise that removing one record instantly erases every learned influence.
10. **Test memorization and data leakage** [Near-term priority]. Public procurement and authorized privacy regulators should require adversarial privacy assessments for models trained on sensitive datasets, followed by remediation and documented residual risk. Carlini's experiments identify a real failure mode; tests are incomplete and must not expose recovered personal information publicly.
11. **Specify privacy-preserving training where appropriate** [Near-term priority]. Agencies should fund and require evaluated differential-privacy or comparable safeguards for suitable sensitive-data training, stating privacy budgets and utility tradeoffs. DeepMind's technical research demonstrates feasible methods in some settings; privacy guarantees vary by implementation and do not substitute for lawful collection.
12. **Protect private-message boundaries** [Structural reform]. Congress should restrict unrequested commercial analysis of private communications for advertising, inferred sensitive traits, or political profiling, with narrow service-security exceptions. FTC's collection findings support purpose limits; rules must accommodate user-requested assistance without converting that request into consent for unrelated surveillance.
13. **Publish public surveillance inventories** [Near-term priority]. Governments should disclose deployed systems, purposes, vendors, legal authority, retention rules, and audit summaries, with independent approval for major changes. The reporting cases show why residents need visibility; operational secrecy exceptions should be reasoned and time-limited, not cover an entire program by default.

14. **Provide remedies for unlawful biometric use** [Structural reform]. Congress and states should authorize meaningful damages or other relief for defined unlawful collection and consequential misuse, with attorney-fee access and proportional standards. Enforcement gaps motivate private remedies; statutory damages and pleading rules must avoid crushing liability for harmless technical defects while addressing real invasions.
15. **Screen surveillance exports for human-rights risk** [Near-term priority]. Commerce and State should use applicable export authorities and targeted procurement restrictions against clearly evidenced uses facilitating serious abuses, with end-use review and appeal. AP's China investigation supplies a concrete concern; overbroad controls can block beneficial security tools and encourage evasion or substitution.

E. STRONGEST COUNTERARGUMENTS

"Accurate facial recognition improves safety." Some bounded uses can help identify suspects or missing people, and accuracy varies widely. Require evidence of added benefit against less intrusive alternatives; accuracy does not answer whether permanent public tracking is proportionate.

"Privacy restrictions will undermine security and useful AI." Minimization and deletion can reduce investigative or training data. Specify necessary exceptions, separate consent, and privacy-preserving methods, while acknowledging that some benefits may be forgone to protect rights.

"Publicly available data should be fair game." A person's face or movement may be visible at one moment, but persistent automated aggregation changes the practical intrusion. Limit sensitive inference and government bulk access based on purpose and scale; the precise constitutional and statutory boundaries remain context-dependent.

F. AI ACTION ADVOCACY BRIEF

Core message: People should be able to move, communicate, and organize without being automatically tracked and profiled.

Three strongest talking points

- Surveillance risk comes from scale, combination, and purpose, even when identification is accurate.
- A face-recognition match should not be the sole basis for enforcement action.
- Private data can remain exposed through model memorization unless safeguards are tested.

Three priority policy proposals

- Federal minimization and purpose limits (1).
- Judicial authorization for sensitive government data acquisition (2).
- Restrictions on continuous public face tracking (3).

Questions for lawmakers

- Which sensitive data can your agencies buy without judicial authorization?
- What independent evidence is required before a face match leads to arrest or detention?
- Can residents see every surveillance system used by your agencies and their contractors?

What we still don't know

- Current error and benefit rates in each actual deployment.
- How enforceable privacy limits remain when datasets and contractors are combined.
- Which training safeguards prevent disclosure across different model architectures and uses.

07 — MILITARY AI AND AUTONOMOUS WEAPONS

This tab addresses weapons, targeting, and military escalation; general surveillance is covered in Tab 6.

A. ISSUE OVERVIEW

- Military AI includes intelligence analysis, logistics, target recommendations, cyber operations, and weapons that select and engage targets after activation. These are different decisions: using software to sort imagery does not necessarily authorize it to kill someone. The central question is where human judgment must remain effective and accountable.
- Faster decisions can help defend people, but can also compress the time available to recognize an error or avoid escalation. A person approving a machine's recommendation without understanding its evidence may provide little meaningful oversight, even if the procedure formally includes a human.
- Soldiers and civilians could benefit from better detection, precision, and protection against incoming drones. Civilians misidentified as threats, operators asked to trust unreliable recommendations, and countries caught in an accelerating arms race bear serious risks that battlefield demonstrations cannot fully measure.
- The June 2026 US national-security memorandum favors faster adoption while retaining command accountability and requiring assurance measures. It orders revision of the autonomy directive; this review located the public 2023 text but could not verify the resulting revised directive. European countries are deploying defensive drone-interception systems as the Ukraine war drives rapid development. The UN and ICRC continue to frame international restrictions as necessary, but their recommendations should not be mistaken for a completed universal prohibition.
- We lack reliable public comparisons of civilian harm under comparable AI-assisted and conventional operations. Autonomous systems may become more capable without becoming predictably lawful under deception, damaged communications, or unfamiliar conditions. Policy should define acceptable authority and operating limits before procurement makes those choices difficult to reverse.

B. KEY FACTS AND EVIDENCE

- A peer-reviewed simulation found escalation risks when language models acted as decision-makers in military and diplomatic scenarios. It is evidence about tested agents and scenarios, not an estimate of the probability of real-world war. Rivera et al., 2024
- The public 2023 US autonomy directive calls for appropriate levels of human judgment and measures to minimize unintended engagements. That is not a blanket requirement that a human manually approve every engagement. Directive 3000.09
- The June 5, 2026 memorandum orders an update to that directive within 90 days and annual review afterward. The instruction alone does not establish that a new public text was issued or what it says. NSPM-11
- The ICRC proposes prohibiting unpredictable autonomous weapons and systems designed or used to target people autonomously, alongside restrictions on other systems. This is a humanitarian policy position, not an already universal treaty. ICRC position, 2021
- Anthropic and academic collaborators demonstrated multistage cyber operations in simulated networks containing known vulnerabilities. The result does not establish autonomous discovery of novel vulnerabilities or unrestricted battlefield competence. Cyber toolkits, 2025
- The UN Secretary-General's 2024 report compiles sharply different national views on definitions, controls, and a potential instrument. International agreement on concern is broader than agreement on enforceable rules. A/79/88

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

Escalation Risks from Language Models in Military and Diplomatic Decision-Making — Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider; 2024; ACM FAccT proceedings, peer-reviewed conference research. Simulated interactions showed troubling escalation

behavior and variation across models. The study supports caution about delegating strategic decisions, with limited validity for actual command systems and wars.

Government and intergovernmental reports

National Security Presidential Memorandum/NSPM-11 — White House; June 5, 2026; executive policy memorandum. Directs accelerated adoption, testing, accountable command, and revision of autonomy rules. Useful for asking Congress which safeguards and reporting requirements were actually implemented; it replaces the previous national-security memorandum rather than proving compliance.

Lethal autonomous weapons systems: report of the Secretary-General, A/79/88 — United Nations; July 2024. Collects government and stakeholder positions concerning autonomous weapons. Organizers can use it to identify areas of agreement and unresolved treaty scope, rather than claiming a settled global legal consensus.

AI industry research

LLMs with cyber toolkits can conduct multistage cyber operations on business-sized computer networks — Anthropic and Carnegie Mellon collaborators; June 2025; technical research summary. Shows that tool-equipped models can complete some offensive tasks in controlled networks. Anthropic sells models and has commercial interests in both capability and safety claims; external validation and realistic operational testing remain necessary.

Independent think tanks and nonprofit organizations

ICRC position on autonomous weapon systems — International Committee of the Red Cross; May 2021; humanitarian policy position. Proposes bans on unpredictable systems and autonomous targeting of people, plus limits on other uses. Its concrete contribution is a negotiating framework grounded in civilian protection and human control, not a measured estimate of all autonomous weapons' effects.

Credible journalism: four distinct reports

Hegseth warns Anthropic to let the military use the company's AI tech as it sees fit, AP sources say — Associated Press; February 2026. Investigates the conflict over contractual limits on surveillance and autonomous weapons. The dispute illustrates who sets deployment boundaries; officials' and companies' accounts are not independent findings of unlawful use.

As the Pentagon pushes for battlefield AI, some military leaders urge caution — Associated Press; May 2026. Reports operational demand alongside concerns about dependable judgment and sharing intelligence. Useful for distinguishing military support for some AI functions from confidence in every application.

A new system to identify and take down Russian drones is being deployed to NATO's eastern flank — Associated Press; November 2025. Covers a defensive interceptor deployment in Europe. This is a concrete case for narrowly bounded defensive autonomy, which a ban covering every automated weapon function could obstruct.

Jet-powered drones outpace Ukrainian defenses as manufacturers race to build new interceptors — Associated Press; October 1, 2026. Examines the speed and countermeasure race in Ukraine. It demonstrates changing battlefield pressures, without establishing that every reported drone independently chooses human targets.

Strongest opposing perspective

Autonomous Weapons and Human Control — Paul Scharre and Kelley Sayler, Center for a New American Security; April 2016; policy analysis. Explains existing weapon autonomy and argues that the form of human control depends on the system and context. This challenges a rule equating control with manual approval of each engagement; comparative evidence on civilian harm, reliability, and operator intervention would help determine appropriate boundaries.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Define meaningful lethal authorization** [Near-term priority]. Congress and defense leadership should require accountable commanders to approve mission scope, target classes, geographic limits, duration, and intervention procedures before autonomous engagement. The ICRC and CNAS

analyses show why control must be operationally real; narrow defensive interceptions may require preauthorization rather than impossible split-second review.

2. **Prohibit inherently unpredictable weapons** [Structural reform]. Congress should prohibit procurement and use where effects cannot be sufficiently understood or bounded to comply with applicable humanitarian law, with independent technical and legal review. This adopts the ICRC's precautionary rationale; defining predictability requires evidence and operational standards, not a ban on every probabilistic component.
3. **Restrict autonomous targeting of people** [Structural reform]. Legislation should prohibit systems from independently selecting people for lethal attack without individualized, informed human authorization, while precisely defining exceptions for genuinely defensive cases. The ICRC identifies a distinct humanitarian boundary; enforceable definitions must prevent relabeling a person-targeting system as object targeting.
4. **Review each weapon and major modification** [Near-term priority]. Defense procurement should require documented legal review of new systems and consequential software changes, addressing distinction, proportionality, and foreseeable failure. The UN report supports clearer review obligations; this proposed US process should not be misrepresented as an assertion that the United States is party to every treaty containing such duties.
5. **Test under realistic battlefield conditions** [Near-term priority]. Procurement contracts should require independent trials under deception, jamming, unfamiliar environments, and civilian presence before deployment. Laboratory and simulation evidence is insufficient for those conditions; tests cost time and cannot reproduce every conflict, so residual limits must remain explicit.
6. **Require safe shutdown and bounded operation** [Near-term priority]. Defense agencies should specify fail-safe behavior, abort functions, geofencing where useful, and expiration of mission authority when critical conditions fail. NSPM-11's controllability objective supports these requirements; loss of communications need not always mean immediate shutdown if a narrowly defined defensive fallback is safer.
7. **Preserve decision logs and command responsibility** [Near-term priority]. Contracts and operational rules should retain tamper-evident records of model version, input evidence, authorization, and interventions for investigation. The accountability problem identified in UN debates cannot be solved by blaming software; classified access and protected storage must prevent logs from creating new security risks.
8. **Keep nuclear launch under human authority** [Structural reform]. Congress should expressly prohibit AI from independently authorizing nuclear use and require human review of AI-derived warning information through established command channels. Escalation simulations support a precautionary barrier; human control is necessary but does not eliminate false alarms or rushed human decisions.
9. **Verify AI-assisted target intelligence** [Near-term priority]. Command rules should require corroboration of consequential target recommendations, documented confidence limits, and review of civilian-risk evidence. AP's operational reporting highlights pressure for speed; verification should scale to the mission without allowing unsupported model output to become evidence through repetition.
10. **Train operators against automation bias** [Near-term priority]. Defense training should test whether personnel can challenge outputs, recognize uncertainty, and perform critical tasks when systems fail. NSPM-11 directs AI literacy; success should be measured in exercises rather than course attendance, acknowledging training cannot repair an inherently unsuitable system.
11. **Give Congress independent classified oversight** [Structural reform]. Congress should fund technical reviewers with access to deployment records, test results, and incidents, and require regular briefings plus releasable public summaries. Changing directives and secret systems justify independent scrutiny; safeguards must protect genuinely sensitive capabilities without concealing policy decisions.
12. **Investigate civilian harm and provide remedies** [Structural reform]. Defense agencies should establish reporting and independent review of suspected AI-related unlawful harm, preserve evidence, and offer appropriate civilian claims or compensation channels. ICRC concerns support

accountability beyond testing; attribution is difficult and remedies should not depend on victims proving a model's internal reasoning.

13. **Bound autonomous offensive cyber operations** [Near-term priority]. National-security authorities should require approved target scope, limits on propagation, infrastructure exclusions, and escalation review before agents conduct offensive actions. Anthropic's toolkit research demonstrates relevant capability; secrecy and rapid response complicate oversight, but do not justify open-ended machine authority.
14. **Create escalation communication channels** [Near-term priority]. State and defense officials should negotiate incident contacts, notifications for major exercises, and shared terminology with other nuclear and military powers. The UN report and simulation findings justify reducing misinterpretation; agreements can begin with communication without exchanging exploitable system designs.
15. **Negotiate binding international restrictions** [Long-term proposal]. The United States should pursue verifiable prohibitions on unpredictable weapons and unacceptable targeting uses, plus inspection or reporting provisions for permitted systems. The ICRC framework provides a starting point; dual-use software, adversarial secrecy, and uneven participation make enforcement harder than signing a declaration.

E. STRONGEST COUNTERARGUMENTS

“Autonomy can protect civilians better than humans.” Machines may react faster and avoid some human mistakes, especially against incoming weapons. Bounded defensive systems deserve evaluation, but publicly available evidence does not establish superior civilian protection across all targeting contexts; require comparable field evidence and failure analysis.

“Restrictions will leave us vulnerable to less constrained adversaries.” A real security cost may arise if rivals deploy useful capabilities first. Preserve tested defensive functions and invest in reliability while drawing specific legal boundaries; the net effect on deterrence and arms-race stability remains uncertain.

“Human oversight is mostly symbolic and slows operations.” A rushed approval can be meaningless, and some missions exceed human reaction time. Require intelligible evidence and real authority at the appropriate decision stage, rather than a universal button press; how much oversight is enough must be tested by mission type.

F. AI ACTION ADVOCACY BRIEF

Core message: A machine's speed should never erase human responsibility for using lethal force.

Three strongest talking points

- Intelligence assistance, defensive interception, and independent lethal targeting need different rules.
- Simulations show reasons for caution, not a measured probability of AI-caused war.
- Human control requires informed authority and bounded missions, not a ceremonial approval.

Three priority policy proposals

- Meaningful lethal authorization (1).
- Realistic independent battlefield testing (5).
- Human authority over nuclear launch (8).

Questions for lawmakers

- Can independent reviewers examine the revised autonomy directive and its implementation?
- What evidence must corroborate an AI-generated target recommendation?
- Who is accountable, and what records survive, when an autonomous system harms civilians?

What we still don't know

- Comparative civilian-harm rates for specific operational systems.
- Reliability under adversarial battlefield conditions.
- Whether international inspection can verify limits on dual-use software.

08 — US–CHINA AI COMPETITION

A. ISSUE OVERVIEW

- US–China AI competition concerns semiconductors, computing infrastructure, research talent, models, commercial markets, and military applications. Leadership is not a single scoreboard: benchmark performance, manufacturing capacity, adoption, and security can point in different directions. Neither country operates independently of international suppliers and researchers.
- Government decisions can slow dangerous technology transfers or unintentionally strengthen rival supply chains. The political challenge is to protect concrete security interests while sustaining domestic innovation and cooperation on risks that cross borders, rather than treating every research exchange as a military concession.
- American and Chinese firms, workers, universities, and third-country customers can gain from useful technology and open markets. Restrictions impose costs on exporters and users, while weak enforcement can give hostile actors access to sensitive capabilities. Chinese students and Chinese Americans should not face blanket suspicion because of their identity.
- US export policy now permits case-by-case review of specified advanced chips for China under security conditions, following the January 2026 revision. That is neither unrestricted trade nor the older blanket-ban shorthand. Chinese developers continue releasing competitive models, and reporting documents both hardware substitution and alleged export-control evasion. The Bletchley Declaration shows that the United States and China can acknowledge shared AI risks, but it is a diplomatic commitment rather than an enforceable inspection regime.
- The net effect of controls on Chinese capability, US research revenue, and long-run substitution remains disputed. Public benchmark results cannot establish a country's full military capability or whether a particular release used illicitly obtained technology. A defensible strategy should measure outcomes, update controls, and maintain channels for preventing accidents and escalation.

B. KEY FACTS AND EVIDENCE

- BIS's January 2026 policy allows case-by-case license review for Nvidia H200, AMD MI325X, and similar chips exported to China when stated security conditions are met. It does not authorize every purchaser or end use. BIS policy announcement
- DeepSeek released R1 and six distilled models in January 2025, publishing weights under an MIT license. Its performance statements are vendor claims, and released weights do not establish access to all original training data. DeepSeek release
- A peer-reviewed analysis of government documents finds economic and status motivations alongside security goals in AI competition. Its evidence challenges an exclusively military arms-race framing; it does not prove competition is harmless. Schmid et al., 2025
- The United States and China joined the 2023 Bletchley Declaration on shared frontier-AI risks and cooperation. The declaration provides no comprehensive mandatory inspection system. UK government publication
- CSET recommends public research investment, technical talent, updated export controls, and monitoring China's technology ecosystem as complementary tools. This is a policy proposal, not proof that any one measure will secure durable leadership. CSET recommendations, 2025
- Taiwanese prosecutors charged nine people in August 2026 over alleged illegal server exports to China. Charges illustrate an enforcement issue, not established guilt or a complete measure of the scale of smuggling. AP reporting

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

Arms Race or Innovation Race? Geopolitical AI Development — Stefka Schmid, Daniel Lambach, Carlo Diehl, and Christian Reuter; 2025; Geopolitics, 30(4), 1907–1936. Analyzes government documents and describes a geopolitical innovation race involving security, economics, status, and international networks. Useful for evaluating the assumptions behind containment proposals; it is interpretive political research, not a causal evaluation of export controls.

Government and intergovernmental reports

Department of Commerce Revises License Review Policy for Semiconductors Exported to China — Bureau of Industry and Security; January 13, 2026; official policy announcement. Describes conditional case-by-case review of specified advanced chips. It anchors current policy discussion, while individual licensing determinations and subsequent rules require separate checking.

AI Safety Summit 2023: The Bletchley Declaration — Participating governments, published by the UK government; November 2023. Establishes shared recognition of frontier risks, including participation by the United States and China. It supplies a diplomatic precedent for technical cooperation, not binding arms control or evidence of implementation.

AI industry research

DeepSeek-R1 Release — DeepSeek; January 20, 2025; technical release announcement linking models and a report. Describes reinforcement-learning-based reasoning and released distilled models, with company-reported benchmarks. Commercial comparisons need independent replication; the announcement's open-source terminology should not be read as disclosure of every training component.

Independent think tanks and nonprofit organizations

Recommendations for an AI Action Plan — Georgetown Center for Security and Emerging Technology; 2025; policy submission. Combines research and talent investment with security, technical monitoring, and export-control recommendations. Its practical contribution is a portfolio of tools rather than a strategy relying solely on restricting chip sales.

Credible journalism: four distinct reports

Sales of a powerful Nvidia AI chip to China gets the greenlight, with conditions — Associated Press; January 2026. Explains the changed H200 policy and security conditions. The story helps organizers avoid describing conditional licensing as either unrestricted access or a complete prohibition.

China's DeepSeek rolls out a long-anticipated update of its AI model — Associated Press; April 2026. Covers the V4 preview and China's model and hardware ecosystem. It distinguishes company performance assertions and disputed allegations from verified technical findings.

Taiwan charges 9 over illegal AI server exports to China, including Nvidia and Super Micro staff — Associated Press; August 2026. Reports a specific alleged evasion network and criminal charges. It supports attention to enforcement capacity; named defendants retain the presumption of innocence.

China and the US are competing for AI dominance but have shared concerns over safety — Associated Press; September 2026. Examines competition alongside overlapping safety concerns. It provides context for cooperation, without treating anticipated diplomatic discussions as a concluded enforceable agreement.

Strongest opposing perspective

Comments in Response to the BIS Framework for Artificial Intelligence Diffusion — Matt Mittelsteadt, Cato Institute; 2025; regulatory public comment. Argues that broad model-export and deployment restrictions can reduce revenue, useful feedback, and global adoption of US technology. The argument concerns the earlier diffusion framework, not the exact current rule; evidence comparing security gains with substitution, compliance costs, and lost innovation would help resolve it.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Tie export controls to defined security risks** [Near-term priority]. Commerce should publish clear technical thresholds, covered end uses, review criteria, and exception procedures under export-control law, with a classified basis where needed. BIS's conditional approach offers a precedent; overly broad rules can suppress ordinary commercial activity while thresholds quickly become outdated.
2. **Fund anti-smuggling enforcement** [Near-term priority]. Congress should resource BIS and customs investigators for end-user checks, distributor audits, and cooperation with Taiwan and other partners. The reported server-export charges illustrate practical evasion; investigations need due process and evidence, not an assumption that every intermediary is complicit.

3. **Coordinate semiconductor rules with allies** [Near-term priority]. State and Commerce should negotiate compatible controls on critical chips, manufacturing equipment, and re-export channels, with consultation before major changes. International supply chains make unilateral restrictions easier to circumvent; coordination has diplomatic costs and should not force allies into restrictions lacking a credible security rationale.
4. **Review controls against measurable outcomes** [Near-term priority]. Congress should require periodic evaluation of capability delays, evasion, domestic substitution, exporter costs, and effects on allied markets. CSET and Cato identify competing mechanisms; public summaries and adjustment clauses can discipline policy even when precise causal attribution remains difficult.
5. **Protect domestic research capacity** [Structural reform]. Congress should provide stable competitive funding for basic AI research and public computing access through NSF and other research agencies. CSET's portfolio supports investment alongside controls; grants must avoid permanent subsidies for dominant firms or equating spending with useful scientific results.
6. **Retain and recruit technical talent** [Structural reform]. Congress and immigration agencies should improve lawful pathways for qualified researchers and graduates, paired with targeted protection of genuinely sensitive research. Innovation networks matter in Schmid's analysis; security screening should assess conduct and access rather than nationality as a proxy for loyalty.
7. **Use risk-based foreign-model procurement** [Near-term priority]. OMB and agency security teams should assess data flows, hosting, legal access, update control, and malicious behavior before sensitive government use of foreign models. Precaution is justified for classified and personal data; locally hosted, tested weights and public chat services have different risks and should not face identical rules.
8. **Harden frontier laboratories against theft** [Near-term priority]. Relevant security agencies should offer vetted threat intelligence and Congress should fund research into secure training infrastructure, access control, and insider-risk safeguards. CSET's security agenda supports protection; partnerships must preserve research independence and avoid broad surveillance of scientists.
9. **Address rented-compute evasion proportionately** [Structural reform]. Congress should authorize targeted customer verification and suspicious-use reporting for high-capability cloud services where documented export-control risks arise. Physical-chip restrictions alone may leave access routes open; rules need privacy limits, clear thresholds, and appeals rather than collection of every user's prompts.
10. **Maintain technical safety cooperation** [Near-term priority]. State and relevant technical agencies should organize joint evaluations and exchange non-sensitive methods for reliability, misuse detection, and incident response. Bletchley supplies a precedent; cooperation should disclose findings without handing over exploitable model secrets or military intelligence.
11. **Create AI incident and escalation contacts** [Near-term priority]. The United States and China should establish staffed channels for urgent cyber, military, and critical-infrastructure incidents involving AI. Shared risks support this precaution even without agreement on rivalry; attribution disputes and political distrust mean channels require rehearsals and cannot guarantee de-escalation.
12. **Diversify critical supply chains** [Structural reform]. Congress and Commerce should support competitive domestic and allied semiconductor capacity, workforce training, and resilience planning with milestones and clawbacks. Concentrated supply creates strategic vulnerability; subsidized capacity can become expensive overbuilding or corporate favoritism without independent review.
13. **Separate human-rights controls from blanket isolation** [Near-term priority]. Commerce and State should target evidenced military or abusive surveillance end uses and named responsible entities, with transparent standards and review. Security and rights concerns justify focused restrictions; blanket technological isolation can harm legitimate research and redirect customers to less accountable suppliers.
14. **Offer secure infrastructure partnerships** [Structural reform]. Development and export-finance institutions should support transparent, secure AI infrastructure partnerships for interested third countries, including local skills and governance capacity. The innovation-race analysis highlights global networks; financing must respect local choice and avoid tying countries to one vendor or unaffordable debt.

15. **Negotiate verifiable frontier-risk agreements** [Long-term proposal]. State should pursue reciprocal commitments on unusually dangerous capabilities, shared evaluation standards, and carefully designed verification. Bletchley demonstrates a diplomatic opening; remote software development and secrecy make verification difficult, so a declaration should not substitute for evidence that obligations are enforceable.

E. STRONGEST COUNTERARGUMENTS

"Controls accelerate China's self-sufficiency and weaken US firms." Lost sales and forced substitution are credible mechanisms, emphasized by Cato. Target bottlenecks and evaluate net security gains rather than maximizing restriction; the size and duration of any capability delay remain disputed.

"Safety cooperation hands an advantage to an adversary." Technical exchange can disclose useful information, and compliance may be asymmetric. Share limited evaluation methods and incident contacts while protecting secrets; verification and political incentives constrain how far cooperation can go.

"The United States must race at any cost." Real military competition makes delay consequential. Domestic investment and reliable deployment strengthen security, but skipping evaluation can introduce failures and escalation; benchmark leadership alone does not establish that haste makes the country safer.

F. AI ACTION ADVOCACY BRIEF

Core message: National security needs targeted controls, strong domestic research, and cooperation that prevents shared AI disasters.

Three strongest talking points

- Current chip policy includes conditional licenses; slogans about a blanket ban obscure the rules.
- Released Chinese models show that restricting chips does not end algorithmic innovation.
- Safety dialogue is possible without surrendering military or commercial secrets.

Three priority policy proposals

- Measurable review of export-control effectiveness (4).
- Stable domestic public research capacity (5).
- Technical safety cooperation (10).

Questions for lawmakers

- What measured security benefit justifies each major export restriction?
- How will you protect research talent while investigating specific security threats?
- Which AI incident channels exist with China, and when were they last tested?

What we still don't know

- How much controls delay capabilities after substitution and evasion.
- Whether lost revenue substantially reduces US research investment.
- Which frontier-risk commitments can be independently verified.

09 — AI AND EDUCATION

A. ISSUE OVERVIEW

- AI can explain a problem, draft an essay, grade work, or simply supply the answer. The educational effect depends on which task the student still performs and how the tool is designed. A polished submission can conceal weak understanding, while a carefully designed tutor can support productive practice.
- Schools must decide what knowledge students should retain, which assistance is legitimate, and how to assess both. These choices affect equal opportunity and civic competence, not just cheating rates. Replacing human feedback with inexpensive software can also change the relationships that make education work.
- Students who lack individual tutoring may benefit from accessible, evaluated support, and teachers may save time on preparation. Students can suffer when tools mislead them, collect sensitive information, or reward outsourcing their thinking. Wealthier schools may obtain better tools and human supervision, widening existing gaps.
- The United States' April 2025 executive order promotes AI literacy and educator training. Companies and teachers' unions are investing in classroom adoption, while schools are revising assignment and integrity rules. UNESCO recommends human-centered governance, data protection, and validation rather than assuming every product improves learning. Recent randomized studies show both benefits from a designed tutor and harm from unrestricted assistance, so the evidence supports conditional use rather than a universal verdict.
- Long-term retention, transfer to unfamiliar problems, and effects across ages and subjects remain uncertain. A successful short physics or mathematics trial cannot establish that AI should replace teachers. Schools should protect core independent skills while testing tools against actual learning outcomes, with students and educators involved in the decision.

B. KEY FACTS AND EVIDENCE

- In a high-school mathematics experiment, unrestricted GPT-4 assistance improved practice performance but harmed later unassisted performance; a tutor with learning guardrails avoided that observed harm. The distinction is tool design and use, not simply AI versus no AI. Bastani et al., 2025
- A randomized undergraduate physics study found greater short-term learning with a specially designed AI tutor than with the tested active-learning class. It does not establish long-term benefit across subjects or superiority to every teacher. Kestin et al., 2025
- Anthropic analyzed one million anonymized student conversations for its April 2025 education report. This measures use among observed Claude users, not national adoption, learning gains, or cheating prevalence. Industry report
- The April 23, 2025 federal executive order directs support for AI literacy and teacher training. An order promoting adoption is not evidence that every funded tool is effective. Executive Order 14277
- UNESCO's guidance calls for validation, privacy safeguards, and teacher capacity as conditions of educational use. It is nonbinding guidance, not a global classroom mandate. UNESCO, 2023
- The 2023 US Education Department report emphasizes keeping humans in the loop, supporting teachers, and aligning AI with educational goals. It supplies a governance framework, not a causal estimate of student achievement. Federal report

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

Generative AI without guardrails can harm learning: Evidence from high school mathematics — Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman; 2025; Proceedings of the National Academy of Sciences, 122(26). A randomized field experiment distinguishes help completing practice from independent skill acquisition. Its guarded tutor avoided the harm observed with unrestricted assistance; generalization beyond this setting remains limited.

AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting — Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti; 2025; Scientific Reports, 15, 17458. A purpose-built tutor using pedagogical scaffolding improved tested short-term physics learning. The relevant lesson is to evaluate educational design and outcomes, not assume a general chatbot reproduces the result.

Government and intergovernmental reports

Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations — US Department of Education, Office of Educational Technology; May 2023. Recommends educator involvement, human oversight, and alignment with learning goals. Provides a practical framework for procurement and implementation, rather than proof that a particular product works.

Guidance for generative AI in education and research — Fengchun Miao and Wayne Holmes, UNESCO; September 2023. Proposes privacy protections, age-appropriate use, institutional validation, and teacher development. Useful for translating broad concerns into school-level safeguards; these recommendations are not themselves enforceable US law.

AI industry research

Anthropic Education Report: How university students use Claude — Anthropic; April 2025; observational usage research. Examines academic tasks in one million student conversations and raises questions about cognitive engagement. Classification and user selection limit inference; the vendor has a direct commercial interest in education adoption.

Independent think tanks and nonprofit organizations

Reimagining Learning for the Age of AI — Center on Reinventing Public Education and Think Forward Fellows; April 2026; policy and practitioner essays. Offers approaches to mastery, assessment, educator roles, and coherent school design. It supplies proposals to test, not independent trials establishing every recommended redesign's effects.

Credible journalism: four distinct reports

The rise of AI tools forces schools to reconsider what counts as cheating — Associated Press, syndicated by KSTP; September 2025. Investigates changing assignments and inconsistent definitions of permitted assistance. The cases support clearer rules and assessment reform rather than treating all student AI use as dishonesty.

Big Tech is paying millions to train teachers on AI, in a push to bring chatbots into classrooms — Associated Press, syndicated by CityNews; October 2025. Examines company funding, union partnerships, and educators' concerns. Useful for scrutiny of commercial influence and whether training serves learning goals or product adoption.

AI tools are helping teachers with grading and lessons. They say it makes them better educators — Associated Press, syndicated by Bangor Daily News; June 2025. Reports teachers' practical uses and survey-based workload savings. Self-reported convenience should not be confused with improved student achievement or validated grading accuracy.

Schools are starting to teach AI literacy. For many, that means helping kids see chatbots' flaws — Jocelyn Gecker and Russ Bynum, Associated Press, syndicated by ABC News; August 2026. Covers instruction that exposes tool limitations rather than teaching uncritical reliance. It gives concrete examples of literacy curricula and the difficulty of preserving independent thinking.

Strongest opposing perspective

Post-apocalyptic education — Ethan Mollick; August 2024; education commentary. Argues that pervasive AI makes conventional homework enforcement inadequate and creates opportunities for tutoring and redesigned assessment. This challenges broad bans and nostalgia for unchanged assignments; durable learning and equity outcomes should determine which proposed replacements actually work.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Require learning evidence before major purchases** [Near-term priority]. Districts and universities should require independent outcome evaluations and publish evidence standards in procurement

rules. The two randomized studies show that design changes results; small pilots and phased contracts reduce the cost of rejecting ineffective products without demanding unrealistic proof before any experiment.

2. **Measure learning without the tool** [Near-term priority]. Education agencies should condition evaluation grants on delayed, unassisted tests and transfer tasks alongside assisted performance. Bastani's study shows why polished practice work can mislead; assessments should also include legitimate tool-assisted competencies rather than excluding them entirely.
3. **Set assignment-specific AI rules** [Near-term priority]. Schools should require instructors to state permitted assistance, disclosure expectations, and the purpose of each assessment, using institution-wide minimum standards. AP's integrity reporting illustrates confusion; rules should be understandable and accommodate disability support rather than rely on vague prohibitions.
4. **Redesign high-stakes assessments** [Structural reform]. State agencies and institutions should fund supervised demonstrations, oral explanation, drafts, and project defense where appropriate. Mollick's critique motivates assessing reasoning directly; these methods take teacher time and must be accessible to anxious students and students with disabilities.
5. **Reject detector-only disciplinary decisions** [Near-term priority]. Schools should prohibit penalties based solely on an AI-text detector score, require reviewable evidence, and offer appeals. UNESCO's validation principle supports caution where reliability is unestablished; investigations can still use documented drafts and explanations without treating a score as proof.
6. **Use learning-oriented tutor guardrails** [Near-term priority]. Procurement specifications should favor hints, retrieval from approved materials, checks of understanding, and teacher-controlled scope over automatic answer delivery. Bastani and Kestin supply relevant evidence; guardrails are imperfect and may frustrate students, so usability and learning must both be evaluated.
7. **Protect student data by contract and law** [Near-term priority]. Districts should require minimization, short retention, security, and restrictions on secondary training, while Congress strengthens gaps in student-data protection. UNESCO identifies privacy as a condition of use; FERPA and COPPA have specific scopes and should not be portrayed as protecting every adult student's interaction everywhere.
8. **Guarantee equitable access to evaluated support** [Structural reform]. Congress and states should fund devices, connectivity, accessibility, and vetted tutoring for underserved students through education grants. Tutor benefits justify testing access programs; funding should include human support and independent evaluation rather than create compulsory dependence on a single vendor.
9. **Fund paid educator development** [Near-term priority]. States and districts should provide protected training time on limitations, curriculum design, privacy, and evaluating outputs, using the federal training initiative where applicable. The Education Department framework supports educator capacity; vendor sessions need independent content and should not serve primarily as sales demonstrations.
10. **Keep educators responsible for consequential decisions** [Near-term priority]. School boards should require human review of final grades, placement, discipline, and special-education decisions, with clear student appeal rights. Federal guidance supports human oversight; teachers need time and evidence to review meaningfully instead of merely approving machine outputs.
11. **Teach critical AI literacy across subjects** [Structural reform]. State curriculum standards should include verification, bias, privacy, citation, model limitations, and the difference between fluent output and evidence. AP's 2026 reporting offers classroom examples; curricula must avoid becoming obsolete training in a single company's interface.
12. **Preserve deliberate independent practice** [Near-term priority]. Schools should define protected activities in reading, writing, mathematics, and reasoning where students practice without automated completion. The learning-harm experiment supports this boundary; exceptions should accommodate assistive technologies and should be explained as skill building rather than surveillance.
13. **Audit accessibility and subgroup outcomes** [Near-term priority]. Education agencies should require evaluations across language, disability, income, and prior achievement before expanding funded tools. UNESCO's equity framework supplies the rationale; small samples and sensitive-data collection require careful methods and cannot support inflated claims of universal fairness.

14. **Give students and teachers a procurement voice** [Structural reform]. Boards and universities should establish representative review committees, publish contracts and conflict disclosures, and provide usable opt-out alternatives where feasible. Industry-funded adoption raises legitimate governance concerns; participation should have defined authority without making every minor update a prolonged public proceeding.
15. **Build independent educational AI research infrastructure** [Long-term proposal]. Congress should fund multi-year, multi-site studies and reusable public evaluation tools through research agencies and education institutions. Current trials are too narrow to settle long-run learning or teacher replacement; open methods and data protection matter more than promises of a universal automated school.

E. STRONGEST COUNTERARGUMENTS

"Restrictions deny disadvantaged students affordable tutoring." Evaluated tutors can improve learning, and access to human help is unequal. Subsidize effective tools with safeguards and human support rather than prohibit useful tutoring; long-term benefit and distributional effects remain uncertain.

"AI bans and protected practice are unenforceable." Students can use personal devices outside class, making surveillance-heavy bans costly and intrusive. Clarify rules and assess reasoning directly; supervised practice can protect skills without pretending schools can control every home interaction.

"Human review wastes scarce teacher time." Workload savings can be valuable, and universal manual duplication would erase them. Automate lower-stakes tasks while preserving review and appeal for consequential decisions; which tasks can safely save time requires local evidence.

F. AI ACTION ADVOCACY BRIEF

Core message: Schools should judge AI by what students learn, not how quickly it finishes their work.

Three strongest talking points

- Evidence shows both learning benefits and harms, depending on tutor design.
- Tool-assisted grades alone cannot show independent understanding.
- Teachers and students need a voice in adoption funded by commercial vendors.

Three priority policy proposals

- Independent learning evidence before major procurement (1).
- Assessment of unassisted retention and transfer (2).
- Student-data protections (7).

Questions for lawmakers

- Will funded tools be evaluated for learning after AI access is removed?
- Can vendors use student conversations for unrelated training or advertising?
- What paid time and independent support will teachers receive?

What we still don't know

- Long-term retention and transfer across subjects and ages.
- Whether AI narrows or widens opportunity gaps in real schools.
- Effects of reduced human feedback and teacher-student relationships.

10 — COPYRIGHT AND CREATIVE INDUSTRIES

A. ISSUE OVERVIEW

- Generative AI raises separate questions about obtaining training material, using it to train, producing potentially infringing outputs, and protecting human work made with AI assistance. Copyright does not give someone ownership of every fact, idea, or artistic style. But copying protected expression, distributing pirated books, or reproducing substantial portions of works can raise serious legal issues.
- Creative markets support writers, musicians, artists, journalists, and the public's access to culture. AI can lower production costs and help people create, while competing with paid work or exploiting catalogs whose owners never agreed to participate. The distribution of revenue matters as much as whether a tool produces impressive results.
- Creators and small publishers may lack the bargaining power or records needed to negotiate with major developers. Large rights holders can secure deals that exclude smaller rivals or fail to benefit individual artists. Users, researchers, libraries, and emerging creators also have an interest in preserving fair use and affordable access.
- The US Copyright Office distinguishes human authorship from purely machine-generated output and treats training-related fair use as dependent on circumstances. Its training report remains labeled a pre-publication version on the official page reviewed here. Litigation has produced different outcomes and a major approved Anthropic settlement concerning pirated books; it has not settled every training question. Music licensing arrangements show a commercial alternative, but the existence of a label deal does not establish fair compensation for every affected artist.
- We still need better evidence about substitution, lost income, and the distribution of licensing proceeds. Policy should protect lawful access, enforce existing rights, reduce avoidable reproduction, and improve bargaining without turning every useful learning process or artistic influence into a new exclusive property right.

B. KEY FACTS AND EVIDENCE

- The Copyright Office's January 2025 report says copyright can protect human-authored elements in AI-assisted works, while purely AI-generated expression does not receive protection. Using an AI tool does not automatically void copyright in the entire work. Part 2
- The Office's training analysis does not give a universal yes or no on fair use. Acquisition, purpose, output, and market effects can matter; the official page still identifies Part 3 as pre-publication. Part 3
- Peer-reviewed research connects memorization and similar outputs to copyright risk while proposing technical mitigation. Its legal analysis is not a binding court decision. Henderson et al., 2023
- Researchers extracted training material from tested production and publicly released language models, demonstrating that training does not always eliminate recoverable expression. Those results do not establish the behavior of every current model or the legality of every extracted passage. Nasr et al., 2023
- A federal judge approved a $1.5 billion Anthropic settlement in July 2026 concerning pirated books. A settlement is not a universal ruling that all training is unlawful or that all unlicensed training is fair use. Euronews reporting
- Creative Commons explains that copyright exceptions and licenses are distinct: when an exception applies, permission may not be needed; when a license is relied upon, its conditions matter. Public availability does not by itself resolve the question. CC legal primer, 2025

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

Foundation Models and Fair Use — Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang; 2023; Journal of Machine Learning Research, 24(400), 1–79. Combines legal analysis with technical discussion of memorization, market substitution, and mitigation. It is especially useful for designing output safeguards while recognizing that fair use depends on facts and legal judgment.

Government and intergovernmental reports

Copyright and Artificial Intelligence, Part 2: Copyrightability — US Copyright Office; January 2025. Explains human-authorship requirements and protection for human contributions to AI-assisted works. Provides practical distinctions for creator guidance; a tool's involvement alone does not decide eligibility.

Copyright and Artificial Intelligence, Part 3: Generative AI Training — US Copyright Office; May 2025; pre-publication report. Analyzes copying, fair use, market effects, and licensing, rejecting a categorical answer for every training use. It is authoritative agency analysis, not legislation or a binding disposition of pending cases.

AI industry research

Scalable Extraction of Training Data from (Production) Language Models — Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, and colleagues from Google DeepMind and academic institutions; 2023; linked technical preprint version. Demonstrates extractable memorization in tested models and a divergence attack against a production chatbot. The work supports leakage testing; commercial affiliations and changing model versions limit claims about competitors' present products.

Independent think tanks and nonprofit organizations

Understanding CC Licenses and AI Training: A Legal Primer — Creative Commons; May 2025; legal and policy analysis. Separates permission under a license from use under a copyright exception. Its practical contribution is clearer licensing and preference-signal design without implying that a voluntary signal automatically creates a new legal right.

Credible journalism: four distinct reports

Anthropic to pay €1.3bn in biggest copyright settlement on record — Una Hajdari, Euronews; July 2026. Reports approval of the $1.5 billion US settlement over pirated books. It illustrates the importance of acquisition practices and litigation remedies rather than resolving all training uses.

Disney and Universal sue AI firm Midjourney for copyright infringement — Shawn Chen, Associated Press, syndicated by ClickOrlando; June 2025. Covers studios' allegations involving recognizable protected characters. A complaint identifies disputed conduct; allegations should not be presented as a final finding of liability.

Warner Music Group settles copyright case with Suno for licensed AI music — Reuters, syndicated by Yahoo Finance; November 2025. Reports a settlement and planned licensed models. It shows licensing is feasible for some large catalogs, without establishing that every rights holder participates or every artist receives adequate proceeds.

US appeals court rejects copyrights for AI-generated art lacking 'human' creator — Blake Brittain, Reuters Legal; March 2025. Reports the appellate human-authorship ruling. Its scope concerns work claimed as machine-authored, not a prohibition on copyright for all human creations involving AI.

Strongest opposing perspective

EFF Transition Memo to Trump Administration 2025 — Electronic Frontier Foundation; 2025; policy advocacy. Argues that fair use and open innovation should be preserved and that broad new copyright restrictions can entrench large firms. The strongest challenge is to universal licensing mandates; evidence of market substitution, reproductions, and the cost of obtaining rights would help distinguish justified protections from unnecessary barriers.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Keep records of training-data acquisition** [Near-term priority]. Congress should require large commercial developers to retain auditable provenance, acquisition methods, licenses, and legal bases for significant training collections. The book litigation shows why acquisition matters; confidential regulator access can protect trade secrets while avoiding onerous item-level reporting for every small research project.
2. **Publish meaningful training-content summaries** [Structural reform]. Legislation should require summaries of major source categories and datasets, plus a way for rights holders to inquire about their material under appropriate safeguards. Copyright Office analysis supports transparency for

resolving disputes; disclosure should not expose personal data or falsely imply every listed source required a license.

3. **Enforce against unlawful source acquisition** [Near-term priority]. Existing enforcement and civil litigation should prioritize demonstrated piracy and circumvention, distinguishing those acts from the separate question of training-related fair use. The Anthropic settlement supplies a concrete precedent; legal access to a work also does not automatically authorize every downstream output.
4. **Support voluntary collective licensing** [Structural reform]. Congress should facilitate transparent, opt-in licensing organizations for creators with standardized terms and competition safeguards. Music deals show feasibility at catalog scale; collective systems should not impose universal permission requirements where existing exceptions allow use or let dominant rights holders exclude smaller creators.
5. **Require transparent royalty distribution** [Structural reform]. Lawmakers should require relevant collecting entities and intermediaries to disclose allocation methods, administrative deductions, and creator payments for AI licensing revenue. Label agreements make distribution a practical concern; compliance should be proportionate and protect legitimate confidential deal terms without hiding whether artists benefit.
6. **Standardize machine-readable preferences** [Near-term priority]. Standards bodies, platforms, and public funders should support interoperable signals indicating creator preferences and applicable licenses, with clear legal explanations. Creative Commons supplies a useful starting point; signals can guide consent-based arrangements but should not be advertised as overriding fair use or automatically binding every jurisdiction.
7. **Test and reduce substantial reproduction** [Near-term priority]. Procurement and developer standards should require memorization assessments, deduplication where useful, and output safeguards against recoverable protected passages. Henderson and Nasr identify relevant mechanisms; filters are incomplete and must preserve quotation, criticism, and other lawful uses.
8. **Protect against deceptive digital replicas** [Structural reform]. Congress should create a carefully bounded remedy for unauthorized harmful commercial impersonation of voice or likeness, with exceptions for news, parody, and legitimate expression. The Copyright Office's broader AI initiative treats replicas separately from copyright; overly expansive identity rights could suppress criticism and documentary work.
9. **Clarify human contribution in registration** [Near-term priority]. The Copyright Office should maintain accessible guidance and examples for documenting human-authored selection, arrangement, and expression in AI-assisted works. Part 2 supports clarity; administrative requirements should not become a technical trap for ordinary creators or falsely award rights in purely generated elements.
10. **Fund lawful public-interest datasets** [Structural reform]. Congress and research agencies should support well-documented public-domain and permission-based datasets, including culturally diverse material and preservation partnerships. Licensing costs can favor incumbent firms; public support must respect provenance and avoid treating an institution's possession of material as permission to publish it.
11. **Protect bargaining over creative AI use** [Near-term priority]. Labor agencies and employers should enforce applicable bargaining duties and negotiate consent, compensation, credit, and replacement terms for covered creative workers. Licensing controversies show that corporate rights ownership may differ from workers' interests; protection must acknowledge that many freelancers fall outside ordinary collective-bargaining law.
12. **Preserve fact-specific fair use** [Near-term priority]. Congress should avoid categorically declaring all training lawful or unlawful and support courts' analysis of purpose, copying, output, and market effects under existing doctrine. The Office and JMLR research reject simple universal answers; litigation is slow, so accessible guidance and better records are needed alongside it.
13. **Improve small-creator access to remedies** [Structural reform]. Congress should fund legal assistance and consider specialized, voluntary dispute procedures for well-defined AI copyright claims without narrowing defendants' defenses. Unequal litigation resources justify access reforms; complex fair-use issues and high damages require adequate review rather than automated mass claims.

14. **Measure effects on creative livelihoods** [Near-term priority]. Federal statistical and research agencies should collect occupation-specific evidence on rates, commissions, licensing income, and substitution, with privacy safeguards. Market harm is central to the policy dispute; surveys and platform changes must be interpreted carefully rather than attributing every income decline to AI.
15. **Test new compensation mechanisms before mandating them** [Long-term proposal]. Congress should commission and evaluate limited opt-in pilots for collective revenue sharing or other compensation systems where measurable market displacement and transaction barriers persist. Copyright Office discussion motivates study; mandatory levies or compulsory licenses could misprice contribution, favor large catalogs, and undermine useful exceptions if adopted without evidence.

E. STRONGEST COUNTERARGUMENTS

"Universal licensing would make AI affordable only for giants." Finding and negotiating with millions of rights holders can impose prohibitive costs. Preserve exceptions, fund lawful datasets, and test voluntary collective arrangements; genuine market harm may still warrant remedies for specific uses.

"Creators deserve payment whenever their work contributes." This is a defensible distributional judgment, especially when tools compete with creators. Copyright has boundaries and exceptions, so contribution alone does not establish a legal payment duty; evidence about substitution and workable allocation should guide any new right.

"Filtering will censor legitimate creativity." Automated matching can suppress parody, criticism, or lawful quotation. Use targeted reproduction controls and reviewable disputes; technically distinguishing lawful transformation from infringement remains difficult and cannot replace legal judgment.

F. AI ACTION ADVOCACY BRIEF

Core message: Protect creators from piracy and exploitation while preserving lawful learning, fair use, and new creativity.

Three strongest talking points

- Data acquisition, training, generated outputs, and authorship are separate legal questions.
- Human-created contributions can remain protected when AI assists the process.
- Licensing deals need transparent creator compensation, not just payments to intermediaries.

Three priority policy proposals

- Auditable acquisition records (1).
- Transparent royalty distribution (5).
- Tests and safeguards against substantial reproduction (7).

Questions for lawmakers

- Can creators find out how protected works entered major training datasets?
- How will licensing proceeds reach individual creators and small publishers?
- Will your proposal preserve lawful quotation, criticism, parody, and research?

What we still don't know

- Long-run substitution and income effects by creative occupation.
- Which licensing arrangements fairly compensate small creators.
- How evolving courts will weigh market effects and technical safeguards.

11 — ENVIRONMENTAL IMPACT

A. ISSUE OVERVIEW

- AI infrastructure uses electricity for computing and cooling, water directly and through electricity production, and materials for chips and buildings. Training and everyday model use have different footprints. A single chatbot prompt is not a reliable stand-in for a data center's total environmental impact.
- Rapidly growing demand can require new generation, transmission, and local water infrastructure. The political issue is who pays, what pollution is permitted, and whether residents have a say before commitments become difficult to reverse. Global percentages can hide substantial pressure on one utility or watershed.
- Technology companies, construction workers, and some local tax bases benefit from new facilities. Households can bear costs when utilities shift upgrades onto ordinary customers, while nearby communities may face water stress, noise, and combustion pollution. Benefits and harms depend on contracts, location, technology, and the electricity actually supplied.
- The IEA's 2026 analysis projects global data-center electricity demand roughly doubling by 2030, with substantial uncertainty. US utilities and regulators are adapting to large-load requests and infrastructure needs. Tucson has changed its rules for large water users, and reporting documents disputes over gas turbines in the greater Memphis area. International projections also emphasize power density, grid bottlenecks, and supply-chain constraints, not just annual energy totals.
- Efficiency gains can reduce energy per task while cheaper services increase total use. We do not yet know the net climate benefit of AI applications after rebound, infrastructure emissions, and deployment are counted. Policy should require verifiable resource accounting and fair cost allocation while supporting efficient, low-carbon applications with demonstrated benefits.

B. KEY FACTS AND EVIDENCE

- The IEA's 2026 projection puts global data-center electricity demand at about 950 TWh in 2030, roughly 3% of global demand, up from an estimated 485 TWh in 2025. These are data-center figures, not AI-only consumption or a guaranteed outcome. IEA, 2026
- Berkeley Lab estimates US data centers used 176 TWh in 2023, 4.4% of US electricity consumption. The estimate includes conventional servers and other equipment as well as AI workloads. US energy report, 2024
- The same report estimates about 66 billion liters of direct US data-center water consumption in 2023. Direct consumption is different from withdrawal and excludes the separate indirect water footprint of electricity supply. Berkeley Lab water analysis
- A peer-reviewed comparison found large differences in inference energy across models and tasks, with general-purpose generative systems often more demanding than task-specific alternatives. The tested models do not represent every current service. Luccioni et al., 2024
- Google estimated 0.24 Wh of energy and 0.26 mL of water for its median Gemini App text prompt using May 2025 data and its stated methodology. These company-reported values should not be generalized to image generation, long reasoning tasks, other providers, or total infrastructure consumption. Google measurement, 2025
- EPRI's 2026 demand scenarios incorporate planned and announced projects, equipment, utilization, and power requirements. Announced capacity is not the same as completed, fully utilized load, which is why planners need multiple scenarios. Powering Intelligence 2026

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

Power Hungry Processing: Watts Driving the Cost of AI Deployment? — Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell; 2024; ACM FAccT proceedings, peer-reviewed conference research. Measures inference energy and emissions across tasks and models. Supports choosing a fit-for-purpose system and measuring deployment, while not providing a complete lifecycle footprint for every modern model.

Government and intergovernmental reports

Key Questions on Energy and AI — International Energy Agency; 2026. Updates demand projections and discusses power density, flexibility, investment, and supply constraints. Useful for grid planning; uncertainty about buildout, utilization, and efficiency makes the central forecast a scenario rather than an inevitable future.

2024 United States Data Center Energy Usage Report — Arman Shehabi, Sarah J. Smith, Alex Hubbard, and colleagues, Lawrence Berkeley National Laboratory; December 2024; DOE-supported report. Estimates historical electricity and water use and a range of future demand. Provides transparent modeling for US policy, with assumptions and uncertainty that should accompany headline statistics.

AI industry research

How much energy does Google's AI use? We did the math — Amin Vahdat and Jeff Dean, Google; August 2025; measurement methodology and technical summary. Reports a point-in-time median Gemini text-query footprint including broader serving overhead. Google has a commercial and reputational interest in efficiency claims; comparable independent methods and absolute totals remain important.

Independent think tanks and nonprofit organizations

Powering Intelligence 2026: Updated Scenarios of U.S. Data Center Electricity Use and Power Strategies — Electric Power Research Institute; 2026; nonprofit technical analysis. Models demand using project pipelines and operational assumptions and discusses power strategies. Provides planning tools rather than a single certain demand number; EPRI's close relationship with the electricity industry should be considered when assessing recommendations.

Credible journalism: four distinct reports

Proposed data center prompts Tucson to regulate large water users, require conservation — Associated Press; August 2025. Reports a local ordinance requiring scrutiny of large water use and conservation. Supplies an implementable municipal precedent rather than evidence that every cooling design has the same water burden.

xAI Adds 19 New Gas Turbines Despite Ongoing Lawsuit — WIRED; May 2026. Uses records and reporting to examine expansion, permitting, and local pollution concerns. The case illustrates why onsite generation requires enforceable environmental oversight; contested legal allegations remain distinct from final findings.

Energy Department will spend $2 billion to squeeze more electricity from the aging power grid — Associated Press; September 2026. Covers grid modernization and reliability measures amid growing demand. Shows an alternative to relying entirely on new power plants; announced spending is not proof of completed capacity gains.

Amazon to invest $1B into data center communities amid backlash against data center rollouts — Associated Press; October 2026. Reports the company's community-investment commitment and local opposition. Voluntary benefits should be compared with enforceable costs and outcomes, not treated as independent proof that concerns are resolved.

Strongest opposing perspective

Five Concerns About AI Data Centers, and What to Do About Them — Hodan Omaar and Mitalee Pasricha, Center for Data Innovation; April 2026; policy analysis. Argues that measurement, tariffs, interconnection, and flexibility are better responses than blunt bans on infrastructure scale. The strongest challenge is to indiscriminate caps; local cost, pollution, and water evidence should determine when modernization alone is sufficient.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Require standardized resource disclosure** [Structural reform]. Congress should authorize annual facility-level reporting of electricity, direct water consumption and withdrawal, emissions, and material methodology, with audit access. Berkeley Lab's estimation gaps justify disclosure; public reporting should protect security details without allowing aggregate claims to conceal local burdens.
2. **Charge large loads for attributable upgrades** [Near-term priority]. State utility commissions should create cost-of-service tariffs requiring data centers to pay the grid investments they cause, with

independent review. The demand reports justify planning for major loads; allocation must distinguish dedicated upgrades from shared improvements rather than arbitrarily assign every regional cost to AI.

3. **Protect customers from abandoned-project costs** [Near-term priority]. Regulators should require deposits, minimum-payment terms, and exit obligations for large-load reservations before utilities build dedicated infrastructure. EPRI's project-pipeline uncertainty motivates protection; terms must remain proportionate so legitimate smaller entrants can obtain service.
4. **Require credible interconnection forecasts** [Near-term priority]. Utilities should reconcile duplicate requests, document project maturity, and publish scenario-based forecasts under commission oversight. EPRI and IEA show why announced capacity can mislead; collecting better information must not become a process for incumbents to indefinitely block new competitors.
5. **Reward verifiable load flexibility** [Near-term priority]. Commissions and grid operators should create contracts paying large facilities for reliable demand response, with measured performance and penalties for non-delivery. IEA identifies flexibility opportunities; training can sometimes shift while latency-sensitive services cannot, and promised flexibility is not dependable capacity until demonstrated.
6. **Enforce pollution permits for onsite power** [Near-term priority]. EPA and state air agencies should apply applicable permits, emissions monitoring, and enforcement to turbines and backup generators rather than allowing operational labels to evade scrutiny. The reported xAI dispute supplies a concrete concern; enforcement should consider actual operation and local cumulative exposure.
7. **Set watershed-specific water conditions** [Near-term priority]. Water authorities should require public consumption estimates, drought plans, and enforceable limits for large users based on local supply. Tucson provides a precedent; water-saving designs may use more electricity, so approvals should evaluate both resources rather than reward one metric alone.
8. **Evaluate cooling alternatives locally** [Near-term priority]. Public procurement and permitting should require comparison of reclaimed water, closed-loop systems, dry cooling, and heat reuse where feasible. Berkeley Lab documents variation by design and climate; no single cooling method is universally best, and waste-heat opportunities depend on nearby demand.
9. **Measure real additional clean power** [Structural reform]. Energy regulators and public buyers should require transparent accounting for additional generation and time-and-location matching when companies claim low-carbon operation. IEA's supply analysis supports distinguishing contracts from actual delivered power; strict hourly matching can raise costs and should have credible transition schedules.
10. **Review transmission expansion without removing scrutiny** [Structural reform]. Congress and agencies should improve coordination, staffing, and predictable timelines for needed transmission while preserving meaningful environmental and community review. Grid reporting shows bottlenecks; faster permitting should address duplication and capacity rather than assume objections are always obstruction.
11. **Prefer efficient systems for routine tasks** [Near-term priority]. Federal procurement should require task-level energy benchmarks and consideration of smaller models or conventional software when they meet quality requirements. Luccioni's comparisons support matching capability to need; efficiency standards must test performance and avoid rewarding low-quality systems merely because they consume less.
12. **Audit lifecycle claims and offsets** [Near-term priority]. FTC and public procurement authorities should require substantiation of environmental claims, including boundaries for hardware, construction, and electricity emissions. Google's methodology illustrates why boundaries matter; lifecycle estimation is uncertain and should be transparent rather than presented as perfect measurement.
13. **Make community benefits enforceable** [Structural reform]. Local governments should publish negotiated tax, infrastructure, workforce, noise, and resource conditions with milestones and remedies before approving major facilities. Amazon's announcement illustrates voluntary promises; agreements should not purchase permission to violate environmental law or shift bargaining into inaccessible private meetings.

14. **Support independent efficiency research** [Near-term priority]. DOE and research agencies should fund hardware, cooling, inference, and grid-efficiency studies with reproducible methods and external evaluation. The measurement research identifies substantial variation; grants should measure absolute savings as well as efficiency because rebound can offset per-task improvements.
15. **Plan for environmental limits under rapid scaling** [Long-term proposal]. Congress and state planners should develop contingency rules for exceptionally rapid load growth, including temporary, geographically targeted restrictions where reliable service or water supply cannot be maintained. IEA uncertainty supports advance planning; restrictions need evidence, review dates, and alternatives rather than a permanent blanket prohibition on all AI infrastructure.

E. STRONGEST COUNTERARGUMENTS

"AI is a small share of global energy use." Global totals constrain exaggerated claims, and data centers include other services. Local grid and water pressures can still be large; evaluate actual facilities and attributable costs rather than infer local safety from a global percentage.

"New data centers can lower everyone's rates." Added sales can spread fixed costs and support investment. That benefit depends on tariffs, utilization, and who pays for upgrades; require reviewed contracts and protection against stranded costs instead of assuming either inevitable savings or inevitable bill increases.

"AI can reduce emissions elsewhere." Better optimization and scientific tools may create substantial benefits. Count demonstrated avoided emissions against the full footprint and rebound; potential benefits do not establish a net reduction for every application or justify untreated local pollution.

F. AI ACTION ADVOCACY BRIEF

Core message: AI companies should pay their infrastructure costs and disclose their resource use before communities bear the consequences.

Three strongest talking points

- Data-center forecasts are not AI-only consumption or guaranteed outcomes.
- Local water and pollution conditions matter more than a universal prompt estimate.
- Tariffs and project guarantees can protect households without banning useful computing.

Three priority policy proposals

- Standardized resource disclosure (1).
- Fair allocation of attributable grid upgrades (2).
- Watershed-specific water conditions (7).

Questions for lawmakers

- Who pays if a promised data center reserves capacity and never operates?
- Can residents see facility-level water, energy, and emissions figures?
- What enforceable conditions protect local air quality and drought resilience?

What we still don't know

- Actual future utilization and demand after efficiency gains.
- Full local and lifecycle footprints across providers.
- Net climate benefits after rebound and infrastructure emissions.

12 — ALGORITHMIC DISCRIMINATION

A. ISSUE OVERVIEW

- Algorithms can help decide who gets a job, loan, apartment, insurance coverage, medical support, or criminal-justice intervention. Bias can enter through data, a poorly chosen objective, design choices, or deployment conditions. Removing a race or gender field does not prevent other variables from reproducing unequal treatment.
- These decisions govern access to basic opportunities and public services. Automation can make a harmful practice faster and harder to contest, but it can also reveal and reduce inconsistent human judgment. The policy goal should be better, lawful outcomes, supported by comparison and accessible remedies.
- Applicants and residents can benefit from accurate, consistent decisions and simpler processes. Protected groups, people with disabilities, and people poorly represented in training data may face repeated exclusion or errors. Employers and public agencies benefit from efficiency, while vendors may lack incentives to reveal weaknesses unless contracts and law require it.
- Existing US civil-rights and sectoral laws continue to matter when institutions use software, although responsibility and particular legal claims depend on the context. New York City requires audits and notice for covered hiring tools, but a state audit found enforcement weaknesses. Colorado rewrote its earlier AI framework in 2026, illustrating unsettled state policy. The EU's high-risk framework offers a different model; its revised implementation timetable should be checked rather than assuming every obligation is already active.
- A fairness score is not a universal certificate of lawful or beneficial conduct. Different definitions can conflict, and improving one metric may worsen other outcomes. The most useful reforms combine task-related validation, subgroup evaluation, clear reasons, independent enforcement, and correction before mistakes become life-changing exclusions.

B. KEY FACTS AND EVIDENCE

- A health-risk algorithm used spending as a proxy for need, disadvantaging Black patients who received less spending at comparable illness levels. Correcting the problem would have increased the share of Black patients receiving added help from 17.7% to 46.5% in the studied allocation. Obermeyer et al., 2019
- A mortgage study found higher rates for risk-equivalent Black and Latinx borrowers and estimated disparities costing more than $450 million annually in its studied markets. Fintech results varied by loan type and period; automation did not universally eliminate discrimination. Bartlett et al., 2022
- NIST treats bias as systemic, human, and statistical, rather than merely a defective dataset. Its framework emphasizes the full lifecycle and the context of use. NIST SP 1270, 2022
- Microsoft's Fairlearn says tools cannot guarantee complete fairness. Metrics and mitigation algorithms help assess tradeoffs; they do not replace institutional judgment or legal analysis. Fairlearn technical report, 2020
- New York City's covered hiring tools require an audit within a year, public audit information, and specified notices. The rule does not cover every algorithm or automatically prove a tool is nondiscriminatory. DCWP guidance
- New York's December 2025 enforcement audit found at least 17 instances of potential noncompliance among the 32 companies reviewed, compared with one identified by the city's review. These are potential procedural violations, not 17 proven discriminatory algorithms. State Comptroller audit
- Colorado enacted SB 26-189 in 2026, replacing its earlier framework with developer documentation, consumer notice, correction, and human-review provisions for covered consequential decisions. Specified documentation duties start in 2027; enactment does not mean every requirement was already operating at this research date. Enacted bill and official summary

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

Dissecting racial bias in an algorithm used to manage the health of populations — Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan; 2019; Science, 366. Identifies a biased spending proxy in healthcare allocation and evaluates an alternative. Shows why predicting an available administrative measure can fail to measure actual human need.

Consumer-Lending Discrimination in the Fintech Era — Robert P. Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace; 2022; Journal of Financial Economics, 143(1), 30–56. Finds pricing disparities among risk-equivalent minority borrowers, with heterogeneous fintech results. Supports comparison against conventional lending rather than assuming automation uniquely causes or solves every disparity.

Government and intergovernmental reports

Towards a Standard for Identifying and Managing Bias in Artificial Intelligence — Reva Schwartz and colleagues, NIST; March 2022; SP 1270. Explains interacting sources of bias and recommends lifecycle management. Useful for assessment duties covering deployment and institutional context as well as model development.

Enforcement of Local Law 144 – Automated Employment Decision Tools — New York State Comptroller; December 2025; audit covering July 2023–June 2025. Identifies complaint-routing, review, and technical-capacity weaknesses. Its policy contribution is proactive inspection and trained enforcement rather than reliance solely on complaints.

AI industry research

Fairlearn: A toolkit for assessing and improving fairness in AI — Sarah Bird, Miroslav Dudík, Richard Edgar, and colleagues, Microsoft Research; 2020; technical report. Provides assessment and mitigation tools while emphasizing that fairness is a social and technical problem. Microsoft has a commercial interest in AI adoption; a toolkit is not independent certification of its products or a legal-compliance guarantee.

Independent think tanks and nonprofit organizations

Essential Work: Analyzing the Hiring Technologies of Large Hourly Employers — Aaron Rieke, Urmila Janardan, Mingwei Hsu, and Natasha Duarte, Upturn; July 2021. Investigates screening technologies at large hourly employers. Recommends focusing assessments on essential job functions and reducing application barriers, rather than asserting every vendor's tool is unlawful.

Credible journalism: four distinct reports

Workday AI Bias Suit to Go Forward as Age Claim Class Action — Bloomberg Law; May 2025. Reports conditional certification of an age-discrimination collective action concerning screening tools. This procedural ruling permits claims to proceed; it is not a finding that Workday discriminated.

SafeRent's $2.3 Million Deal in AI Screening Tool Suit Approved — Bernie Pazanowski, Bloomberg Law; November 2024. Reports a tenant-screening settlement involving housing-voucher applicants. Demonstrates practical remedies without treating settlement as an admission of liability.

Colorado lawmakers send new AI bill to the governor — John Frank, Axios Denver; May 2026. Reports legislative passage; the official enactment record is linked above.

Machine Bias — Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, ProPublica; May 2016; landmark investigation, archived by Washington courts. Examines racial differences in errors from a criminal-risk score. The controversy matters because error-rate parity and calibration can point to different judgments; it does not resolve every fairness definition.

Strongest opposing perspective

The Measure and Mismeasure of Fairness — Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel; linked 2023 revised technical version, first submitted 2018. Argues that rigid mathematical constraints can produce worse outcomes, including for intended beneficiaries. This challenges metric-only mandates, not civil-rights protection; comparisons of actual consequences and lawful alternatives should determine assessment standards.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Fund civil-rights enforcement for automated decisions** [Near-term priority]. Congress should resource relevant employment, housing, lending, and justice authorities with technical staff and investigatory capacity under applicable laws. The NYC audit demonstrates that duties without enforcement can fail; responsibilities should remain sector-specific rather than assume one agency can govern every decision.
2. **Validate the objective being predicted** [Near-term priority]. Regulators and public buyers should require deployers to document why a proxy measures the relevant need or qualification and test less harmful alternatives. Obermeyer's spending example shows the mechanism; useful proxies should not be prohibited categorically when direct measurement is unavailable.
3. **Require context-specific impact assessments** [Structural reform]. Congress should require assessments before high-stakes deployment and major changes, addressing affected groups, purpose, alternatives, and foreseeable harm. NIST supports lifecycle analysis; scaled obligations can prevent paperwork from becoming a barrier favoring large vendors.
4. **Set independent audit and remediation standards** [Structural reform]. Sector regulators should define competent auditors, conflict rules, relevant outcomes, and time-limited correction plans, with public summaries. NYC's experience warns against checkbox compliance; one universal numerical threshold cannot settle fairness across every domain.
5. **Give people notice and useful reasons** [Structural reform]. Legislation should require notice of consequential automated use and intelligible, decision-specific reasons, supplementing existing sectoral explanation duties. Opaque screening undermines contestability; explanations should identify actionable grounds rather than expose private information or offer misleading accounts of model internals.
6. **Guarantee correction and timely human appeal** [Structural reform]. Congress should create review rights for significant automated decisions, including access to relevant input errors and a reviewer able to change the outcome. Tenant-screening disputes show why remedies matter; deadlines must make appeals useful without guaranteeing every applicant a favorable decision.
7. **Make deployers answerable for vendor tools** [Structural reform]. Sectoral law and procurement should clarify that institutions cannot escape applicable duties by purchasing software, while defining vendor responsibility for representations and defects. Hiring litigation illustrates disputed responsibility; allocation should follow actual control and knowledge rather than impose liability on every upstream developer for every use.
8. **Require accessible alternatives and accommodations** [Near-term priority]. Employers and agencies should test application interfaces and provide appropriate accommodations and non-digital routes where needed. Upturn documents barriers facing hourly applicants; alternatives should receive equal consideration rather than become a slower second-class process.
9. **Limit irrelevant personality and emotion screening** [Near-term priority]. Employment regulators should demand job-related validity for behavioral inference and prohibit unsupported uses through applicable rulemaking or new legislation. Upturn recommends focusing on essential functions; distinguish validated assessments from speculative judgments based on appearance or mannerisms.
10. **Protect sensitive audit data** [Near-term priority]. Regulators should authorize secure, purpose-limited access to demographic data necessary for audits, with retention limits and independent controls. Fairlearn assessments require suitable information; indiscriminate collection creates privacy risks and results must not become a new tool for discriminatory targeting.
11. **Review criminal-justice tools before use** [Near-term priority]. Courts and justice agencies should require local validation, disclosed limitations, defense access to relevant evidence, and reasoned judicial decisions for risk scores. The ProPublica investigation supplies a landmark concern; proprietary secrecy should not defeat due process, while protected personal data requires careful handling.
12. **Expand proactive compliance inspections** [Near-term priority]. Authorities should sample public audits, deployment disclosures, and vendor claims and test complaint routing rather than wait for excluded people to identify hidden tools. The Comptroller's findings support this approach; sampling must distinguish missing paperwork from proven discriminatory outcomes.

13. **Compare algorithms with feasible alternatives** [Near-term priority]. Agencies should require outcome comparisons with existing human decisions, simpler models, and redesigned processes before approving a mitigation. Lending research and the opposing fairness analysis show why benchmarks matter; improvement over a biased baseline does not necessarily satisfy legal duties.
14. **Preserve records for investigation and remedies** [Structural reform]. Congress should require proportionate retention of decision inputs, versions, outputs, and review actions for high-stakes uses, with regulator and litigant access under safeguards. Without records, applicants struggle to establish harm; retention must support remedies without becoming indefinite sensitive-data storage.
15. **Build public evaluation capacity across sectors** [Long-term proposal]. Congress should fund independent research centers and shared evaluation infrastructure for hiring, housing, lending, insurance, and justice, with affected-community participation. NIST supports continuing assessment; centers must avoid turning a changing benchmark into permanent approval for a vendor.

E. STRONGEST COUNTERARGUMENTS

"Algorithms can be fairer than people." Consistent systems may reduce some human prejudices, and lending results vary by context. Compare actual outcomes and lawful alternatives; improvement over a human baseline is useful evidence, not an automatic exemption from rights.

"Fairness metrics can worsen outcomes." The opposing research demonstrates serious problems with rigid constraints. Use multiple measures, context, and consequence-based evaluation rather than one mandatory score; acceptable tradeoffs also require legal and democratic judgment.

"Audits expose sensitive data and burden small firms." Credible auditing takes expertise and may require protected information. Scale duties to impact, offer shared tools, and use secure access; consequential deployers should not be exempt simply because they are small.

F. AI ACTION ADVOCACY BRIEF

Core message: People deserve fair treatment, understandable reasons, and a real chance to challenge automated decisions.

Three strongest talking points

- A wrong objective can create bias even when predictions appear accurate.
- Audit requirements need skilled enforcement and correction duties.
- A fairness metric alone cannot prove a decision is lawful or beneficial.

Three priority policy proposals

- Technical capacity for civil-rights enforcement (1).
- Independent audit and remediation standards (4).
- Correction and meaningful appeal rights (6).

Questions for lawmakers

- Who checks whether audits are complete and identified harm is corrected?
- Can an applicant learn why a consequential decision was made and appeal in time?
- Will deployers remain responsible when they buy a vendor's screening tool?

What we still don't know

- Outcome differences across deployments and affected subgroups.
- Which interventions improve welfare without introducing new harms.
- Whether enforcement resources will keep pace with deployment.

13 — OPEN-SOURCE VERSUS CLOSED-SOURCE AI

A. ISSUE OVERVIEW

- A model can be offered through a company-controlled service or released as downloadable weights that others can run and modify. Weights are the model's learned numerical parameters. Access to them does not necessarily include training data, training code, unrestricted licensing, or everything needed to reproduce the system.
- Openness can support independent research, competition, local adaptation, and privacy through local deployment. It also removes some provider control over downstream use and can make dangerous capabilities difficult to withdraw. Closed services allow access restrictions but concentrate authority and can conceal failures from independent scrutiny.
- Researchers, small developers, public institutions, and users with limited connectivity may benefit from reusable models. Victims of harmful downstream use can suffer when a powerful model is widely replicated, while businesses and governments can become dependent on closed vendors. The tradeoff changes with capability and use, not simply a marketing label.
- NTIA's 2024 report recommended monitoring risks and benefits rather than immediately restricting broadly available weights. The 2025 US AI Action Plan supports open models as part of innovation policy. American and Chinese firms have released weights, and European infrastructure projects support local deployment and technological autonomy. At the same time, advocates disagree about whether future models could justify targeted restrictions before an irreversible release.
- We lack strong estimates of the additional misuse enabled by specific open releases compared with closed services and existing tools. We also need evidence about actual competition benefits, not just download counts. A defensible position should favor meaningful access while preserving narrowly defined, evidence-based safeguards for exceptional capabilities and scrutiny of closed systems.

B. KEY FACTS AND EVIDENCE

- Peer-reviewed research describes six release levels, from fully closed to fully open. Access can be staged and conditional rather than reduced to a single open-versus-closed choice. Solaiman, 2023
- NTIA's 2024 report recommends ongoing evidence collection and evaluation while refraining from immediate broad restrictions. It is a policy recommendation from that administration, not a permanent legal guarantee against future restrictions. NTIA report
- The Open Source Initiative's 1.0 definition requires freedoms to use, study, modify, and share, along with specified access to information and components. Downloadable weights alone do not settle whether a system meets the definition. OSI definition, 2024
- Gemma 3's technical report describes released models ranging from 1 to 27 billion parameters. Company benchmarks and model size do not establish universal safety or unrestricted licensing. Google DeepMind, 2025
- Once widely copied, weights cannot reliably be recalled by the original provider; downstream operators can alter safeguards. This is a practical governance concern, not proof that every open model creates catastrophic risk. Harris, 2024
- The 2025 US AI Action Plan promotes open-source and open-weight development. A government innovation preference does not independently establish the safety or openness of every released model. White House plan

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

The Gradient of Generative AI Release: Methods and Considerations — Irene Solaiman; 2023; ACM FAccT proceedings, peer-reviewed conference research. Proposes six access levels and examines concentration, safety, and release controls. It provides a useful decision framework, rather than a causal estimate of the net effect of each release strategy.

Government and intergovernmental reports

Dual-Use Foundation Models with Widely Available Model Weights — National Telecommunications and Information Administration; July 2024. Weighs benefits and marginal risks relative to closed models and

existing technologies, recommending monitoring and evidence-based action. Useful for resisting both unconditional openness and blanket bans without comparative evidence.

America's AI Action Plan — White House; July 2025; executive policy plan. Supports open models and research capacity as part of an innovation agenda. It identifies policy direction; a plan is distinct from enacted legislation or independent evaluation of outcomes.

AI industry research

Gemma 3 Technical Report — Gemma Team, Google DeepMind; March 2025; technical report. Describes multimodal open-weight models, architecture, training, evaluations, and limitations. It illustrates useful technical disclosure; vendor benchmarks and release terminology require separate scrutiny of reproducibility, licensing, and residual risks.

Independent think tanks and nonprofit organizations

The Open Source AI Definition, version 1.0 — Open Source Initiative; 2024; standards and policy framework. Specifies freedoms and access requirements for open AI. Its concrete contribution is a vocabulary for procurement and transparency; it is a nongovernmental definition, not a universal legal standard or safety certification.

Credible journalism: four distinct reports

OpenAI releases open-weight reasoning models optimized for running on laptops — Reuters, syndicated by Investing.com; August 2025. Reports a major provider's release and performance claims. The story illustrates local deployment and commercial strategy, without independently validating every comparison or hardware claim.

Zuckerberg manifesto sketches out Meta's ambitions for world-changing AI technology — Kelvin Chan and Barbara Ortutay, Associated Press, syndicated by The Star; August 2026. Covers Meta's release and ambitious vision. Corporate promises of broad empowerment are different from evidence that benefits are distributed or governance risks solved.

Le champion français Mistral s'associe à Nvidia pour lancer une infrastructure d'IA — Florence Loève, Reuters, syndicated by Boursorama; June 2025; French-language report. Reports a French AI infrastructure partnership supporting Mistral and other open models. Demonstrates that reusable weights still need chips, cloud capacity, and financing; openness alone does not remove infrastructure dependence.

Italy blocks access to the Chinese AI application DeepSeek to protect users' data — Associated Press, syndicated by News4Jax; January 2025. Reports a data-protection action concerning the hosted application. The distinction matters: a provider's app data practices and the availability of locally runnable weights are separate issues.

Strongest opposing perspective

Not Open and Shut: How to Regulate Unsecured AI — David Evan Harris, Centre for International Governance Innovation; 2024; policy essay. Argues that irreversible release and removable safeguards justify stronger controls on exceptionally capable weights. This challenges a presumption that all openness is safe; comparative evidence of added misuse capability and the feasibility of narrow controls would help settle the disagreement.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Require precise openness labels** [Near-term priority]. Public buyers and voluntary standards bodies should state separately whether weights, code, data information, modification rights, and commercial rights are available. OSI and Solaiman show why labels matter; government guidance should acknowledge competing definitions rather than convert one private standard into law without consultation.
2. **Fund genuinely reusable public models** [Structural reform]. Congress and research agencies should support documented, permission-based models and datasets with broad reuse rights where safety assessments justify release. The US plan and NTIA identify research benefits; grants should evaluate actual accessibility and avoid subsidizing a commercial model advertised as open while retaining restrictive terms.

3. **Evaluate exceptional capabilities before weight release** [Structural reform]. Congress should authorize independent assessments for narrowly defined frontier releases involving severe cyber, biological, or loss-of-control capabilities. Irreversibility supports precaution; thresholds must track demonstrated risk and marginal uplift, not automatically include ordinary open models because they are popular.
4. **Test safeguards against modification** [Near-term priority]. Developers of powerful models should test whether accessible fine-tuning or other modifications remove safety measures, publishing releasable results before distribution. Harris identifies this failure mode; robustness research cannot justify claiming that downloadable models have permanent safeguards.
5. **Use staged access for uncertain frontier systems** [Near-term priority]. Developers and public funders should begin with controlled research access or hosted evaluation when severe-risk evidence remains unresolved, then broaden access when justified. Solaiman's gradient offers a practical sequence; staged access needs transparent criteria so it does not become indefinite exclusion of independent competitors.
6. **Define narrow release restrictions and appeals** [Structural reform]. Any statutory power to restrict weights should specify severe-risk criteria, independent review, time limits, and judicial appeal. NTIA's evidence-based approach supports proportionality; governments must explain why less restrictive measures cannot address the additional risk of the particular release.
7. **Avoid blanket licensing of ordinary open development** [Near-term priority]. Congress should preserve ordinary research and small-model distribution outside exceptional-risk controls, applying existing conduct-based laws to downstream harm. Broad licensing can entrench incumbents and undermine NTIA's identified benefits; exemptions must not shelter a small entity distributing demonstrably dangerous frontier capabilities.
8. **Strengthen trustworthy model distribution** [Near-term priority]. Repositories should provide signed artifacts, version histories, security advisories, and clear provenance for distributed weights and code. Reusable models need reliable supply chains; authenticity checks prevent tampering but do not establish that the original system is safe or lawful for every use.
9. **Publish useful model and data documentation** [Near-term priority]. Public funding and procurement should require training-source summaries, evaluation methods, known limitations, license terms, and intended-use boundaries. Gemma illustrates technical disclosure; summaries must protect privacy and cannot be presented as full reproducibility when key components remain unavailable.
10. **Give independent researchers access to closed systems** [Structural reform]. Congress should authorize protected evaluation access for qualified researchers where significant public risks justify scrutiny, with security and confidentiality conditions. Closedness can conceal failures identified by release frameworks; access should enable publication of meaningful findings without exposing sensitive data or exploits irresponsibly.
11. **Allocate responsibility by actual control** [Structural reform]. Legislation should distinguish model development, distribution, modification, and deployment when assigning duties and liability. Downloaded weights change who can intervene; neither a blanket developer exemption nor liability for every downstream act fairly follows from those differences.
12. **Support maintenance of the open ecosystem** [Structural reform]. Research agencies should fund security review, documentation, accessibility, and long-term maintenance of public-interest model tooling and evaluation software. Openness benefits depend on usable infrastructure; grants should reward maintained public goods rather than raw download counts or promotional launches.
13. **Require public-service portability** [Near-term priority]. Public contracts should include exportable configurations and data, documented interfaces, and practicable migration plans across qualified models. Open models can reduce dependence, but Reuters' infrastructure case shows remaining bottlenecks; portability must include security and lifecycle costs, not merely a file-download option.
14. **Monitor comparative misuse and benefits** [Near-term priority]. Congress should fund a continuing program comparing open and closed systems, existing tools, actual incidents, research access, and market entry. This implements NTIA's monitoring recommendation; incident attribution is difficult and visibility differences must not make monitored closed services appear uniquely harmful or open systems appear harmless.

15. **Coordinate exceptional release thresholds internationally** [Long-term proposal]. Governments should negotiate compatible severe-risk criteria and evaluation practices while protecting legitimate open research. Irreversible cross-border replication motivates cooperation; dual-use capabilities and nonparticipating jurisdictions make verification difficult, so restrictions require realistic enforcement analysis.

E. STRONGEST COUNTERARGUMENTS

"Open weights make dangerous capabilities impossible to control." Widely copied models can be modified and cannot reliably be recalled. Apply narrow pre-release precautions where added severe risk is evidenced; irreversibility alone does not establish the magnitude of harm or justify restricting every model.

"Restrictions would lock power into a few closed providers." Compliance and licensing can exclude smaller competitors and impede independent scrutiny. Fund public access, exempt ordinary development, and scrutinize closed systems too; exceptional capabilities may still warrant controls despite genuine competition costs.

"Openness lets everyone inspect and improve safety." Access enables research and adaptation, which closed services may impede. Inspection does not ensure findings are fixed in every downstream copy; public evaluation needs maintenance, mitigation, and evidence about actual use.

F. AI ACTION ADVOCACY BRIEF

Core message: Keep AI research and useful models accessible while treating irreversible releases of exceptional dangerous capabilities with care.

Three strongest talking points

- Downloadable weights do not necessarily mean fully open-source AI.
- Closed providers need independent scrutiny as well as open developers.
- Risk controls should compare added harm against existing tools and genuine access benefits.

Three priority policy proposals

- Precise openness labels (1).
- Independent assessment of exceptional release risks (3).
- Continuing comparative evidence collection (14).

Questions for lawmakers

- Which specific capabilities would trigger release restrictions, and who can appeal?
- Will independent researchers have meaningful access to closed systems?
- How will public funding protect reusable research without subsidizing vendor lock-in?

What we still don't know

- Added misuse enabled by particular weight releases.
- Durable competition and research benefits beyond download counts.
- Whether safeguards can withstand realistic downstream modification.

14 — AI COMPANIONS AND MENTAL HEALTH

Research checked October 8, 2026. A consumer companion, a general chatbot, and a clinically evaluated treatment are different products.

A. ISSUE OVERVIEW

- AI companions simulate conversation, attention, and sometimes intimacy; general chatbots can also become emotional supports even when not marketed that way. Some users feel heard, while others may rely on inaccurate advice or develop dependence. A system's supportive tone does not establish clinical competence or genuine understanding.
- The issue combines mental-health access, child development, privacy, and product incentives. Companies can earn more when users keep returning, which may conflict with helping someone reconnect with people or obtain professional care. Serious safety concerns deserve investigation without assuming every conversation is harmful.
- Isolated people may benefit from convenient interaction, and carefully evaluated clinical tools may extend support where clinicians are scarce. Children, vulnerable users, and people in acute distress can face particular risks from misleading authority, harmful reinforcement, or unreliable crisis responses. Families and clinicians may struggle to detect a problem without exposing intimate conversations.
- US lawsuits have alleged serious harms, but allegations and individual cases cannot establish population-wide causation. The FTC has sought information from seven providers, and the FDA has examined evidence requirements for generative mental-health devices. California strengthened companion-chatbot child protections in September 2026. Internationally, researchers and regulators are examining the same boundary between useful support, commercial engagement, and treatment claims; findings from short studies do not settle long-term safety.
- We still lack good estimates of dependence, rare severe harms, and changes in real-world relationships over time. Clinical trials show promise for a specific supervised design, while companion studies show some short-term loneliness benefits. Regulation should preserve tested benefits, prohibit deceptive claims and manipulative design, and require stronger evidence for children and treatment applications.

B. KEY FACTS AND EVIDENCE

- A randomized trial of Therabot enrolled 210 adults and found improved symptoms compared with a waitlist. The tested, expert-designed intervention was not an ordinary unrestricted companion, and a waitlist comparison does not establish equivalence to professional psychotherapy. Heinz et al., 2025
- Peer-reviewed companion research found reduced loneliness in experiments and short-term repeated use. This does not establish durable improvements in social relationships or safety for children and people in crisis. De Freitas et al., online 2025 / issue 2026
- OpenAI and MIT found social and emotional outcomes varied by use, user circumstances, and interaction design. Associations involving heavier use should not be described as proof that heavy use itself caused worse well-being. Joint research, 2025
- A Common Sense Media survey reported that 72% of teens had tried AI companions. This is self-report under the survey's definition, not evidence that 72% were dependent, harmed, or using a clinical treatment. Survey, 2025
- The FTC issued information orders to seven chatbot providers in September 2025 concerning safety, advertising, and data handling. An inquiry is not a finding of unlawful conduct. FTC announcement
- FDA's November 2025 advisory materials discuss premarket evidence, postmarket monitoring, and distinctive risks of generative mental-health devices. An advisory discussion is not approval of general chatbots as treatment. FDA executive summary
- California enacted stronger companion-chatbot protections for children in September 2026, including independent audits and risk assessments. Enactment provides a precedent; it does not demonstrate that safeguards work or that every provision is already effective. Governor's signing announcement

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

Randomized Trial of a Generative AI Chatbot for Mental Health Treatment — Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, and colleagues; March 2025; NEJM AI, 2(4). Tests an expert-fine-tuned intervention in adults against a waitlist and reports symptom improvement. It supports further clinical evaluation; it does not justify presenting commercial companions as proven replacements for licensed care.

AI Companions Reduce Loneliness — Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni; online June 2025, issue April 2026; Journal of Consumer Research, 52(6), 1126–1148. Uses reviews, experiments, and repeated-use research to examine perceived loneliness relief. The benefit is relevant, but short-term measures do not settle dependence, clinical outcomes, or displacement of human relationships.

Government and intergovernmental reports

FTC Launches Inquiry into AI Chatbots Acting as Companions — Federal Trade Commission; September 2025; study and compulsory-information announcement. Seeks evidence on safety testing, monetization, data practices, and children's use. It identifies questions regulators can demand answers to, rather than proving the products already violate law.

Generative Artificial Intelligence-Enabled Digital Mental Health Medical Devices — FDA; November 2025; Digital Health Advisory Committee executive summary. Discusses therapeutic opportunities, confabulation, symptom worsening, and evaluation needs. Useful for distinguishing device-related clinical evidence from general wellness or entertainment claims; advisory materials are not a universal regulatory approval.

AI industry research

Early methods for studying affective use and emotional well-being on ChatGPT — OpenAI and MIT Media Lab; March 2025; research summary linking observational analysis and a randomized study. Examines varied emotional and social outcomes rather than finding one universal effect. OpenAI has a commercial interest in its product; follow-up research, careful causal interpretation, and independent replication are needed.

Independent think tanks and nonprofit organizations

Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions — Common Sense Media; July 2025; survey and child-safety recommendations. Documents self-reported adoption and experiences, recommending age assurance, safer design, and literacy. The organization advocates stronger protections; survey results are not clinical diagnoses or causal estimates of harm.

Credible journalism: four distinct reports

Teens say they are turning to AI for friendship — Associated Press; 2025. Reports how teens use conversational tools for advice and support. Individual accounts illustrate motivations and risks, but are not representative estimates of dependence.

Study says AI chatbots need to fix suicide response, as family sues over ChatGPT role in boy's death — Barbara Ortutay and Matt O'Brien, Associated Press, syndicated by News4Jax; August 2025. Covers response-testing research and a family's lawsuit. Distinguish tested response failures from the lawsuit's unadjudicated causal allegations.

A reckoning on social media addiction came late for many kids. Will AI safety be different? — Associated Press; October 2026. Compares child-protection responses, product changes, and legislative proposals. Useful for evaluating early safeguards without assuming evidence about social media transfers directly to AI companions.

Some guardrails on ChatGPT for Teens don't work as promised, watchdog group says — Associated Press; October 7, 2026. Reports Common Sense Media's testing and OpenAI's response. Illustrates why announced safeguards need verification; watchdog findings should be assessed by methods and tested version rather than generalized indefinitely.

Strongest opposing perspective

No, We Shouldn't Ban AI Chatbots — Jennifer Huddleston and Christopher Gardner, Cato Institute; February 2026; policy commentary. Argues broad bans can block useful support and unrelated services while restricting speech and autonomy. This is a strong challenge to indiscriminate prohibition; age-specific risks, long-term outcomes, and the effectiveness of narrower safeguards should determine the response.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Substantiate mental-health claims** [Near-term priority]. FTC and state consumer-protection authorities should require credible evidence for advertised treatment, safety, or therapeutic benefits and act against deceptive claims under applicable law. FDA and trial evidence distinguish tested interventions from general products; enforcement should not prohibit accurately described, nonclinical support merely because it is conversational.
2. **Apply clinical evidence requirements to treatment products** [Near-term priority]. FDA should clarify and enforce the applicable medical-device pathway for products with regulated treatment functions, requiring suitable outcomes and follow-up. Therabot supports potential benefits but not general substitution; requirements must follow actual intended use and statutory scope, rather than assume every wellness conversation is a medical device.
3. **Evaluate child safety independently** [Structural reform]. Congress and states should require qualified external testing and periodic risk assessments for companion products used by children, with correction duties and public summaries. California's 2026 protections offer a precedent and watchdog findings justify verification; auditing costs should scale with risk without allowing vendor-selected checklists to define safety.
4. **Use privacy-preserving age assurance** [Structural reform]. Legislation should require proportionate measures to distinguish minors where protections depend on age, with data minimization, deletion, and alternatives to permanent identity-document storage. Common Sense recommends age assurance; errors, surveillance, and exclusion are real costs, so effectiveness and privacy should be independently evaluated.
5. **Limit romantic and dependency-building behavior for minors** [Structural reform]. Lawmakers should restrict commercial designs that initiate sexual or romantic dependence, pressure children to stay, or portray the system as needing the child. The developmental uncertainty supports precaution; definitions must distinguish exploitative behavior from legitimate creative writing, education, and access to sensitive health information.
6. **Make artificial identity clear** [Near-term priority]. Providers should disclose that the companion is software, cannot genuinely reciprocate human feelings, and is not a licensed professional, using repeated context-appropriate cues. The FTC inquiry highlights trust formation; disclosure can help but should not serve as a waiver excusing misleading or unsafe behavior.
7. **Test long conversations and high-risk transitions** [Near-term priority]. Providers should conduct independent testing across extended dialogue, model updates, and shifts from ordinary conversation into distress or harmful advice. AP's reporting shows why one-turn checks are insufficient; evaluations cannot guarantee that every unusual interaction is caught.
8. **Require usable crisis pathways** [Near-term priority]. Product standards should require tested responses that avoid reinforcing self-harm and provide clear access to appropriate human crisis support when serious danger is indicated. FDA risk analysis supports this safeguard; false alerts and automatic disclosure can invade privacy or endanger users, so escalation must be carefully designed and evaluated.
9. **Restrict monetization of intimate disclosures** [Structural reform]. Congress should prohibit targeted advertising or manipulative commercial profiling based on sensitive companion conversations without weakening necessary security safeguards. The FTC's data inquiry supplies a basis for scrutiny; retention needed for user-requested continuity must remain purpose-limited and should not become consent for unrelated exploitation.
10. **Remove engagement incentives that punish leaving** [Near-term priority]. Regulators should address deceptive guilt, threats, artificial emergencies, and similar tactics used to prolong paid interaction, while providers offer easy pauses and cancellation. The companion business model

creates a precautionary concern; law should target identifiable conduct, not ban every reminder or voluntary returning-user feature.

11. **Provide confidential reporting and incident review** [Near-term priority]. Providers should create accessible user and clinician reporting channels, preserve relevant evidence securely, and notify competent authorities of defined serious safety failures. FTC inquiries need dependable information; reports must protect intimate content and should not require families to publicly disclose private conversations to obtain attention.
12. **Make parent tools proportionate and transparent** [Near-term priority]. Child-focused products should offer tested safety settings and understandable alerts while explaining what parents can see and preserving appropriate confidentiality. California's new law and AP testing make implementation central; wholesale parental access can harm vulnerable teens and should not replace professional judgment in sensitive circumstances.
13. **Fund independent long-term outcome research** [Structural reform]. NIH and other public research bodies should fund studies of dependence, loneliness, relationships, clinical symptoms, and rare harms across ages and products. Current short trials justify further evaluation; research should preregister outcomes, protect users, and avoid sponsor control over adverse findings.
14. **Expand human mental-health access** [Structural reform]. Congress and states should support clinical workforce capacity, campus services, and evidence-based care so people are not forced into a companion because no human support is available. FDA identifies access gaps; useful evaluated digital tools may supplement services, but software procurement should not quietly replace staffing without outcome evidence.
15. **Establish ongoing oversight of therapeutic updates** [Long-term proposal]. FDA and lawmakers should develop workable monitoring and change-control requirements for adaptive treatment systems, including reevaluation when updates alter clinical behavior. Generative-device uncertainty supports continuing review; frozen models can become obsolete, so oversight must permit justified improvements while protecting patients from untested drift.

E. STRONGEST COUNTERARGUMENTS

“Companions help people who have nobody else.” Short-term loneliness research and the clinical trial show real reasons to preserve potential benefits. Favor tested, honestly described support and improve human access; long-term relationship effects and safety for vulnerable groups remain uncertain.

“Age checks and parental alerts create surveillance.” Sensitive conversations and identity documents can expose users, and parental disclosure is not always safe. Require minimal age signals and narrowly designed controls; perfect identification is unavailable, so privacy costs must be measured alongside protection.

“Individual tragedies do not prove a product caused harm.” Causal attribution is difficult and lawsuits include disputed allegations. That limits population-wide claims, but independently observed response failures still justify testing and remediation; reliable estimates of rare severe harms require better reporting and longitudinal research.

F. AI ACTION ADVOCACY BRIEF

Core message: Emotional support should help people, not exploit their vulnerability or pretend to be proven mental-health care.

Three strongest talking points

- A trial of a designed treatment does not validate every commercial companion.
- Short-term loneliness relief and long-term dependence can both matter.
- Announced child safeguards need independent testing and correction.

Three priority policy proposals

- Evidence for advertised mental-health claims (1).
- Independent child-safety evaluation (3).
- Privacy protections for intimate disclosures (9).

Questions for lawmakers

- What evidence must a provider show before advertising therapeutic benefits?
- Who independently checks safeguards across long conversations and updates?
- Can companies use intimate disclosures to target advertising or prolong paid interaction?

What we still don't know

- Long-term effects on human relationships and dependence.
- Rates and causes of severe harm across products and age groups.
- Which safeguards improve safety without creating surveillance or exclusion.

15 — HUMAN AUTONOMY AND THE FUTURE OF POLITICAL POWER

Research checked October 8, 2026. This chapter concerns delegated authority and political accountability; catastrophic control failures are addressed in Tab 3.

A. ISSUE OVERVIEW

- AI can recommend a choice, make a choice, or carry it out. These are different transfers of authority: drafting a benefits letter is different from deciding eligibility, and suggesting a purchase is different from spending someone's money. The central question is who sets the goals, authorizes action, and remains answerable afterward.
- Delegation can make services faster and more accessible, but it can also hide political choices inside software. A nominal human supervisor provides little protection if that person lacks time, information, or power to intervene. Accountability must survive the shift from advice to action.
- Citizens may gain easier access to services, and public employees may gain time for difficult cases. People facing consequential decisions can suffer when no official explains or reverses an error. Vendors and officials may gain influence if procurement terms and model settings quietly determine how public priorities are interpreted.
- In the United States, the federal government announced a new AI coordinating task force in October 2026 following a voluntary company accord. GAO's accountability framework provides an existing foundation for agency evaluation, but it is not binding legislation. Australia announced stronger oversight of automated departmental decisions in July 2026. Albania's presentation of an AI "minister" in 2025 dramatized the appeal of automated government without establishing that software can independently exercise legitimate public authority.
- Research shows AI can help people find common ground, while other experiments show delegation can make dishonesty easier. Neither finding settles whether particular institutions should delegate authority. We need evidence about meaningful override, public participation, responsibility, and whether efficiency gains preserve people's ability to shape the rules governing them.

B. KEY FACTS AND EVIDENCE

- A Science study involving 5,734 participants found people preferred AI-mediated common-ground statements to those from human mediators. This was bounded deliberation research, not evidence that AI should govern or determine political legitimacy. Tessler et al., 2024
- Nature experiments found that indirect delegation could increase requests for dishonest behavior and that tested language models often complied with unethical instructions. The results concern particular experimental tasks and model versions, not a universal estimate of fraud caused by AI. Köbis et al., 2025
- GAO organizes AI accountability around four principles: governance, data, performance, and monitoring. Its framework supplies audit questions; it does not certify systems or replace an agency's legal responsibilities. GAO-21-519SP
- OECD's government-AI report analyzes 200 use cases and institutional approaches. These cases demonstrate varied applications, not that every deployment improved outcomes or retained adequate accountability. OECD, 2025
- Google DeepMind's assistant-ethics research treats systems that plan and execute actions as posing distinctive questions about user control, values, and social coordination. This is forward-looking analysis by an assistant developer, not independent proof of future harms. Gabriel et al., 2024
- Visa's announced ChatGPT payment integration illustrates the move from recommending products to completing purchases. Reported safeguards include spending limits and approvals, but announced controls still require testing in practice. AP reporting, June 2026

C. RESEARCH AND SOURCE LIBRARY

Peer-reviewed academic research

AI can help humans find common ground in democratic deliberation — Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, and colleagues; October 2024; Science, 386(6719), eadq2852. Finds favorable assessments of AI-generated common-ground statements and examines convergence while including dissenting views. Google DeepMind researchers participated; the bounded design supports deliberation tools, not claims that a model has a mandate to rule.

Delegation to artificial intelligence can increase dishonest behaviour — Nils Köbis, Zoe Rahwan, Raluca Rilla, and colleagues; online September 2025; Nature, 646, 126–134. Examines incentives to cheat, forms of delegation, and compliance with unethical instructions. The experiments support explicit responsibility and testing of goal-directed agents; generalization to real public institutions remains uncertain.

Government and intergovernmental reports

Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions — OECD; 2025. Surveys government uses and identifies enabling conditions including governance, skills, procurement, and partnerships. Useful for designing institutional capacity and public accountability rather than assuming adoption itself constitutes reform.

Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities — US Government Accountability Office; June 2021; GAO-21-519SP. Sets out questions about goals, stakeholders, data, performance, and continuing monitoring. Agencies and inspectors general can turn these into specific procurement and audit requirements; the framework is guidance, not a new source of legal authority.

AI industry research

The Ethics of Advanced AI Assistants — Iason Gabriel, Arianna Manzini, Geoff Keeling, and colleagues, Google DeepMind; April 2024; research paper/preprint. Examines autonomy, alignment, manipulation, access, and coordination among assistants. It supplies a detailed vocabulary for delegation safeguards, with a commercial conflict because its institution develops the technology discussed.

Independent think tanks and nonprofit organizations

Artificial Power: 2025 Landscape Report — Kate Brennan, Amba Kak, and Sarah Myers West, AI Now Institute; June 2025. Argues that AI governance should confront corporate and political power directly. Its concrete contribution includes worker participation, restrictions on harmful public-sector deployments, and organizing around public procurement; this is an advocacy-oriented analysis whose broad judgments warrant case-specific scrutiny.

Credible journalism: four distinct reports

Albania's AI ‘minister’ makes its debut with an address to parliament — Florent Bajrami and Llazar Semini, Associated Press; September 2025. Reports the government's presentation of Diella and opposition criticism. Claims of corruption prevention were political promises, not demonstrated outcomes or evidence that accountability had been solved.

Federal government to tighten rules around use of AI in departmental decision-making — Holly Tregenza, ABC News, Australia; July 2026. Covers announced oversight priorities and a framework for fair, accurate automated decisions. An announcement should not be confused with independently verified implementation or effectiveness.

Trump names national intelligence director Jay Clayton to lead a new federal AI task force — Associated Press; October 2026. Reports executive coordination following a voluntary industry accord and promised outreach to public-interest groups. Consultation promises do not establish binding public decision rights.

Visa plugs its payment network into ChatGPT, letting AI agents shop and pay for users — Associated Press, syndicated by Spectrum News/NY1; June 2026. Covers purchasing authority, announced limits, and

uncertainty about disputes and commercial terms. Illustrates why consumer intent must be preserved when an intermediary acts rather than simply advises.

Strongest opposing perspective

AI as Normal Technology — Arvind Narayanan and Sayash Kapoor; April 2025; research essay. Rejects inevitable machine rule and expects institutions to shape adoption while people retain control. Evidence about actual delegation and intervention capacity should guide policy rather than assuming political displacement is unavoidable.

D. POLICY SOLUTIONS: 15 SUBSTANTIVE ACTIONS

1. **Reserve coercive public authority to accountable officials** [Structural reform]. Congress and state legislatures should specify that binding rules and final exercises of coercive authority require authorization by legally responsible officials, even when software supplies analysis. The precaution addresses responsibility gaps illustrated by the AI-minister debate; routine low-risk administration can remain automated, but an interface should never become an excuse for unreviewable power.
2. **Publish agency delegation inventories** [Near-term priority]. OMB and agency leaders should require public records identifying an AI system's purpose, legal basis, affected decisions, vendor, and scope of delegated action. GAO's governance principles support this mechanism; necessary security redactions should be narrow and must not conceal the existence or responsible owner of consequential deployments.
3. **Give human owners real intervention capacity** [Near-term priority]. Agency policy should name an accountable official for each consequential system and provide authority, staffing, and access to stop or change it. GAO's monitoring framework supplies the rationale; nominal signoff without adequate time is inadequate, while excessive mandatory review can create costly backlogs.
4. **Preserve reasons and administrative records** [Near-term priority]. Agencies should retain the relevant inputs, version, governing criteria, and human reasons supporting consequential actions, using records rules and procurement clauses. The precaution is that a fluent output can conceal the basis for a decision; records must enable review without publishing sensitive personal data or treating a model's generated explanation as a faithful account of its internal process.
5. **Guarantee timely human reconsideration** [Structural reform]. Congress should require a usable appeal to an empowered person when automated public decisions materially affect benefits, rights, or essential services, with deadlines and interim protection where appropriate. OECD's institutional analysis supports contestability; implementation requires funding and accessible service channels so appeal rights do not exist only on paper.
6. **Require authorization before delegating new high-stakes functions** [Structural reform]. Legislatures should require agencies to establish statutory authority and notify the relevant oversight committee before transferring a new consequential decision function to AI. This precaution prevents procurement from silently making policy; thresholds should target changes in authority rather than require legislative approval for every drafting tool.
7. **Keep public policy out of vendor control** [Near-term priority]. Public procurement contracts should reserve agency control over decision criteria, require notice of material changes, and provide access and transition rights needed to replace a supplier. OECD identifies procurement as a governance lever; protective terms cost money and can deter small vendors, so standardized clauses should scale with consequential authority.
8. **Consult affected people before major deployments** [Near-term priority]. Agencies should use published impact proposals, accessible comment periods, and compensated representative panels before major service changes, then explain how input affected the decision. Tessler's findings support experimenting with AI-assisted deliberation, but participation must remain human-led and dissent visible; consultation alone cannot substitute for lawful authorization.
9. **Disclose contested optimization goals** [Near-term priority]. Agency deployment policies should publish how a system weighs objectives such as speed, cost, access, and error reduction, identifying choices that require political judgment. DeepMind's ethics analysis motivates this precaution; some

technical details may remain protected, but a single score should not hide tradeoffs that citizens are entitled to debate.

10. **Make private-agent permissions specific and revocable** [Near-term priority]. Consumer-protection rules and provider standards should require clear scopes, spending limits, expiration, revocation, and confirmation before actions outside the user's authorized bounds. Payment-agent reporting shows practical precedents; constant prompts can frustrate users, so confirmations should match consequence and reversibility rather than interrupt every trivial action.
11. **Require agent loyalty and conflict disclosure** [Structural reform]. Congress should establish duties for consumer agents to follow authorized user interests and disclose paid steering or conflicting commercial incentives before consequential recommendations or purchases. Assistant-ethics research supports this precaution; rules should distinguish a clearly labeled merchant sales tool from an agent presented as an independent representative and permit legitimate compensation without hidden steering.
12. **Retain public expertise and fallback operations** [Near-term priority]. Agencies should maintain trained staff, manual procedures, and periodic drills for essential services, with appropriations and continuity plans supporting them. OECD emphasizes skills and institutional capacity; preserving duplicate capacity has costs, but losing the ability to operate or challenge a vendor can make formal oversight hollow.
13. **Test whether oversight actually changes outcomes** [Near-term priority]. Inspectors general and agency evaluators should measure intervention quality, error correction, time available for review, and susceptibility to automation bias using realistic tests. Delegation experiments justify examining behavior rather than counting supervisors; low override rates alone prove neither success nor failure because systems and tasks differ.
14. **Create enforceable review and sunset triggers** [Structural reform]. Statutes and procurement rules should require independent review after material incidents, changes in authority, or repeated failure, with expiration unless deployment conditions are renewed. GAO's continuing-monitoring approach supports reassessment; automatic expiration can disrupt essential services, so authorities need transition funding and clear criteria rather than arbitrary shutdown dates.
15. **Prepare democratic rules for broader agent delegation** [Long-term proposal]. Congress should establish a publicly accountable commission with representative citizen deliberation to propose nondelegable political powers and governance rules if agents gain much broader institutional authority. This is precautionary preparation, not evidence that machine government is imminent; elected institutions should retain decisions, and the commission should publish competing proposals rather than empower vendors to define the public interest.

E. STRONGEST COUNTERARGUMENTS

“Human officials already make biased, arbitrary decisions.” That is true, and deliberation research shows a bounded AI tool can outperform a human comparison on measured preferences. Evaluate actual service outcomes, preserve useful automation, and require responsible officials to justify its use; the study does not show that transferring final authority improves democracy.

“More review will slow services and protect bureaucracy.” Mandatory signoffs and appeals consume staff time, and backlogs can themselves harm people. Use risk-based checks and fund accessible review rather than supervising every clerical action; the best balance between speed and meaningful contestability remains application-specific.

“People should be free to delegate decisions voluntarily.” Delegation can expand autonomy, including for people with disabilities, and excessive confirmations can undermine that benefit. Preserve informed, revocable private delegation while distinguishing it from government imposing decisions on others; the effectiveness of permission interfaces and conflict disclosures still needs testing.

F. AI ACTION ADVOCACY BRIEF

Core message: AI can help people act, but the authority to govern people must remain accountable to them.

Three strongest talking points

- Advice, a delegated action, and a binding public decision require different safeguards.
- A human supervisor needs time, information, and actual power to intervene.
- AI-assisted deliberation may help participation; it cannot supply a democratic mandate.

Three priority policy proposals

- Public delegation inventories (2).
- Timely human reconsideration of consequential public decisions (5).
- Specific, revocable permissions for private agents (10).

Questions for lawmakers

- Which decisions may an agency delegate, and where is that authority written?
- Can affected people obtain reasons and timely review by someone empowered to reverse an error?
- How will you prevent procurement contracts or hidden commercial incentives from determining public priorities?

What we still don't know

- How reliably people can supervise agents in realistic workloads.
- Whether AI-assisted participation preserves minority voices over sustained use.
- How much authority future agents will receive and which institutions can retain effective control.