AI-Powered Policymaking: Behavioral Nudges and Democratic Accountability*

Michel Chbeir Jana Dagher

In collaboration with Apart Research

Abstract

Governments around the world are increasingly looking to artificial intelligence (AI) to enhance public policy design, including the use of behavioral "nudges" to influence citizen choices. Nudge interventions—gentle pushes in choice architecture that steer people toward beneficial behaviors without coercion—have proven effective in domains from public health to taxation. By leveraging AI's predictive and personalization capabilities, policymakers hope to refine these nudges for greater impact and efficiency. Simultaneously, these AI-driven approaches raise pressing ethical and democratic questions. Are citizens being transparently informed or unduly manipulated by algorithmic policies? How can we ensure accountability and public trust when government decisions are guided by complex AI systems? This paper examines AI-powered policymaking through behavioral nudges, with a focus on the safety risks of algorithmic interventions. While AI promises greater efficiency and personalization, its use in shaping public behavior introduces potential risks, including manipulation, cognitive security threats, and loss of democratic control. This study evaluates governance models (EU AI Act, Singapore AI Framework) to assess whether current safeguards are sufficient to prevent AI from undermining public autonomy and trust

Keywords:
Social Sciences Track Keywords: behavioral nudging in AI, AI governance, policymaking, democratic accountability, AI transparency, cognitive security, EU AI act, public trust, AI and privacy, dark nudging, nudge theory, algorithmic bias

1. Foundational Theories & Background

1.1. Behavioral Nudges in Policymaking

In public policy, a *"nudge"* refers to any aspect of how choices are presented that alters people's behavior in a *predictable* way without forbidding options or significantly changing economic incentives. In other words, nudges reshape the *choice architecture* – the context in which people make decisions – to gently steer them while preserving freedom of choice. Popularized by Thaler and Sunstein (2008), nudges leverage cognitive biases to promote welfare-improving choices, such as automatic pension enrollment or strategic food placement in cafeterias. This approach, termed "libertarian paternalism," balances guidance with individual autonomy by capitalizing on predictable heuristics. Pioneered by the UK's Behavioural Insights Team in 2010, it has since been widely adopted by governments seeking cost-effective, non-coercive means to positively shape public behavior. By exploiting biases like status quo preference or procrastination, nudges make beneficial choices more accessible without eliminating alternatives, offering a delicately crafted middle ground between laissez-faire and strict regulation.

*Research conducted at the Women in AI Safety Hackathon, 2025

1.2. AI's Role in Refining and Optimizing Policy Nudges

Artificial intelligence elevates behavioral nudging from static, one-size-fits-all interventions to *dynamic, adaptive strategies*—often termed "Nudge 2.0" or *"smart nudges."*. By analyzing vast behavioral datasets on individual preferences, AI personalizes interventions in real time, significantly amplifying their impact. For instance, AI can determine whether an individual responds more effectively to SMS reminders or social comparison cues, tailoring the optimal message accordingly. In healthcare, AI-powered decision support tools act as nudges by providing informed recommended "default" diagnoses, which physicians can override, thus retaining human judgment while improving outcomes, consequently augmenting expert decision-making. Beyond personalization, AI transforms nudging into an **adaptive, continuous, self-optimizing** process. Through *"hypernudging,"* a concept introduced by Karen Yeung, AI refines interventions in real time based on user responses. A smart city's traffic system, for example, could dynamically adjust safety alerts or modify communication strategies to enhance compliance. Digital platforms already employ similar AI-driven architectures, subtly shaping user decisions through personalized interfaces and algorithmic recommendations. Governments, too, can leverage AI-driven choice architectures, such as adaptive e-government portals that steer users toward beneficial actions. While AI-enhanced nudging offers unprecedented **precision and scalability**, its very strengths—personalization, ubiquity, and algorithmic opacity—raise profound ethical and governance concerns. The risk of undermining **autonomy, consent, and transparency** necessitates rigorous scrutiny of its broader societal implications.

1.3. Ethical Concerns in AI-Driven Behavior Modification

The integration of AI into behavioral nudging intensifies longstanding ethical concerns, particularly regarding manipulation and autonomy. While traditional nudges may already exploit cognitive biases, critics warn that AI-driven micro-targeted nudging risks becoming "subliminal" *manipulation*, eroding informed consent and trust, as it operates behind the scenes, unlike conventional visible nudges, making it harder for individuals to recognize when their choices are being skewed. **Transparency** thus becomes a non negligible consideration. **Autonomy** is another critical issue: AI's ability to personalize interventions could lead to hypernudging that **subtly yet irresistibly** steers behavior, diminishing as such "individual autonomy and agency," potentially producing "modulated individuals" whose capacity for self-determination is compromised. If governments use AI to shape public opinion or influence voting behavior, it could undermine democracy. **Privacy concerns** arise as AI-driven nudges often rely on extensive personal data, potentially crossing into **intrusive surveillance**. Large-scale behavioral interventions could blur the line between *benevolent guidance* and *coercive social engineering*, with critics pointing to China's social credit initiatives as a cautionary example. *Justice and fairness* also demand scrutiny, as AI nudges can inherit and exacerbate **biases**. If not properly audited, AI systems may disproportionately target or stigmatize certain groups—such as applying more "anti-fraud" nudges to welfare recipients from marginalized communities. Ethical AI governance must prevent discrimination, ensuring that algorithmic decisions do not further exacerbate or amplify existing biases. In sum, while AI enhances the effectiveness of nudging, it also raises ethical risks. Addressing concerns over manipulation, accountability, and fairness will be crucial as governments explore AI-driven behavioral interventions.

2. Key Case Studies & Governance Models

2.1. European Union AI Act: AI's Role in Government Policy

A champion of establishing solid rules and terms for reliable automated systems, the European Union has pioneered the movement with its **EU AI Act**, adopted in 2024 as the world's first comprehensive regulation for trustworthy AI, with significant implications for government policymaking. It categorizes AI systems into **unacceptable, high, limited, and minimal risk levels**, subjecting each to proportional legal obligations. Notably, the Act *prohibits* AI deemed an **unacceptable risk**, specifically targeting systems that employ subliminal or deceptive techniques to *manipulate* behavior and those used for government-led **social scoring**. For AI classified as *high-risk*, the Act mandates strict requirements encompassing **transparency, accuracy, and robust human oversight**. Providers of such systems must ensure conformity assessments and implement thorough risk management protocols. This regulatory initiative complements existing frameworks like the GDPR, emphasizing algorithmic transparency and accountability, ensuring AI is deployed in a manner consistent with **democratic values, human rights, and the public's trust**. Full enforcement is set for 2027, setting a critical precedent for the responsible use of AI in government.

2.2. Singapore's AI Governance Framework: Regulating AI Decision-Making

Singapore's AI governance follows a *principles-based*, voluntary framework, distinct from the EU's regulatory approach. Introduced in 2019 and updated in 2020, the **Model AI Governance Framework** emphasizes two key **principles**: AI decisions should be *explainable, transparent, and fair*, and AI systems should be *human-centric*, focusing on well-being rather than just efficiency. The framework includes practical pillars such as internal governance, risk management, and stakeholder communication, encouraging organizations to establish oversight structures, ensure human involvement in high-risk AI decisions, and maintain transparency with stakeholders, especially in public services. It also introduced **AI Verify**, a toolkit to assess AI systems for fairness and explainability. Cooperative at its core, this soft law approach relies on existing agencies for oversight, *leaving room for innovation* while still ensuring ethical AI deployment through **guidelines and industry collaboration**.

Aspect	EU AI Act (EU)	Model AI Governance Framework (Singapore)
Legal status	Binding regulation (EU law adopted 2024, with enforcement by 2027). Compliance is mandatory across EU member states.	Voluntary framework/guidelines (first issued 2019, 2nd edition 2020). Not legally enforced; relies on self-regulation and sectoral laws.
Regulatory approach	Risk-based classification of AI uses. Defines *unacceptable risk* (banned practices) and *high-risk* systems with strict requirements. Specific prohibitions on manipulative AI and social scoring by governments.	Principles-based and sector-agnostic. Outlines ethical principles (e.g. transparency, fairness, human-centricity) for all AI systems. Provides guidance rather than hard rules, allowing flexible application across industries.
Scope of coverage	Broad – covers AI systems by both private and public sector that affect EU citizens. Many public sector applications (justice, policing, welfare, etc.) fall under high-risk category with obligations. Aims to protect fundamental rights and safety in all uses.	Broad and cross-sector. Intended for private companies and government agencies alike as a best-practice standard. Focuses on ethical deployment of AI in services and operations; no specific exclusion, but implementation is mainly driven by organizations' adoption.
Key requirements	High-risk AI must meet requirements: rigorous risk assessments, quality and bias management, transparency and the "appropriate level of human oversight". Providers must keep documentation (audit logs, technical docs) and undergo conformity assessments. Non-compliance can lead to significant fines.	Guidelines for good practice: e.g. implement internal governance (AI oversight roles, training); ensure explainability and transparency to users; apply risk-based controls (more oversight for higher-impact AI). Use tools like AI Verify for testing AI against ethical principles. Adoption is encouraged through education, pilot projects, and industry recognition rather than legal penalties.
Notable prohibitions/ principles	Bans AI that *"materially distorts"* human behavior through subliminal or manipulative techniques (avoiding covert behavior modification). Bans AI social scoring by public authorities. Emphasizes human rights (aligns with EU values and GDPR). Also requires certain AI (e.g. deep fakes, biometrics) to have transparency notices (limited-risk category).	No outright bans defined in the framework (since it's non-binding), but warns against exploitative use of AI on vulnerable groups and stresses human-centricity (AI should not undermine human dignity or decision-making). Principles like fairness and accountability imply that manipulative or harmful AI uses are unacceptable. The framework implicitly discourages any use of AI that would violate its ethical tenets, even if not phrased as legal prohibitions.
Enforcement & oversight	To be overseen by national supervisory authorities and an EU-wide AI Board. Strong enforcement: fines up to 6% of global turnover for violations (similar to GDPR levels) have been proposed. Accountability is external – organizations will be audited/certified for compliance.	Oversight is decentralized and voluntary. Organizations self-assess their alignment (e.g. using the ISAGO self-assessment guide). No dedicated AI regulator exists; enforcement relies on sectoral regulators and reputational incentives.

Table 1: Comparison of the EU AI Act and Singapore's governance framework. The EU enforces strict regulations and bans manipulative AI, while Singapore emphasizes ethical guidelines and voluntary compliance to promote transparency and accountability.

3. Ethical Risks & Democratic Accountability

3.1. Transparency in AI-Powered Policymaking

Transparency is a cornerstone of accountable AI in policymaking, ensuring algorithmic decisions remain open to scrutiny and fostering public trust. As AI is integrated into governance, the risk of epistemic opacity raises concerns of unaccountable decision-making. The EU mandates **disclosure and contestability**, while Singapore promotes **participatory AI governance** via voluntary frameworks. Transparency in AI policymaking operates across three key dimensions: **disclosure**, ensuring citizens are aware when AI influences decisions; **explanation**, addressed through *explainable AI (XAI)* and the **GDPR's "right to explanation"**; and **validation**, via accountability reports, audits, and regulatory sandboxes to ensure democratic alignment (as seen in Amsterdam and Helsinki). Past failures, such as the UK's *A-level grading* debacle and the Netherlands' *SyRI fraud detection system*, underscore the need for transparency, revealing how opacity exacerbates bias and erodes public confidence. Without transparency, AI risks undermining **procedural legitimacy and public trust**. Effective AI policymaking must balance **innovation with accountability**, ensuring decisions are explainable, contestable, and aligned with human-centric governance.

3.2. Public Trust in AI-Driven Interventions

Public trust is critical for the successful integration of AI in governance, directly impacting the acceptance and effectiveness of AI-driven policies. Absent trust, citizens may **resist** AI applications, **undermine** policy goals, or **lose confidence** in institutions. Global surveys show *rising public concern about AI*, particularly in the U.S., driven by fears of bias, job displacement, and loss of control. This growing unease underscores the challenge governments face in building public confidence. For AI-driven interventions like nudges, trust is a decisive factor. If AI is perceived as manipulative, it risks backlash. Therefore, *transparency, fairness, and ethical oversight* are crucial. Trust is more readily granted to AI systems embedded within established institutions, such as healthcare. Moreover, public acceptance of AI interventions depends on their perceived intent—those seen as serving citizens' interests, like health reminders, are more widely accepted than those viewed as benefiting the government. To foster trust, governments must ensure robust regulation, demonstrate AI efficacy through tangible success, and engage in transparent communication with the public. By aligning AI with citizens' values and involving them in decision-making, governments can build lasting trust—both a fragile pillar and a fleeting catalyst—requiring **continuous transparency, ethical integrity, and dialogue**.

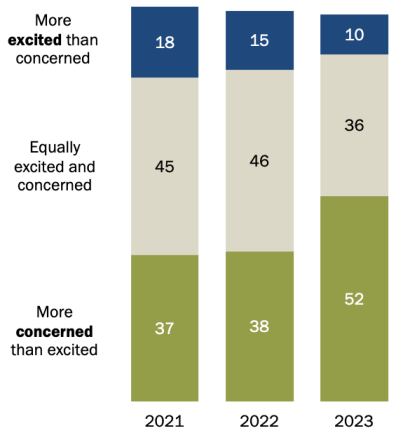

Figure 1: Public concern about artificial intelligence has grown in recent years, far outweighing excitement. In a U.S. survey, **52%** of respondents in 2023 said the increased use of AI in daily life makes them more concerned than excited* (only 10% felt *more excited* than concerned). This reflects a cautious public mood toward AI's expansion. Governments face an uphill task in building trust for AI-driven policies against this backdrop of skepticism.

3.3. Risks of Algorithmic Manipulation in Behavioral Nudges

The ethical risks of AI-driven nudging lie in the potential for **algorithmic manipulation**, which refers to influencing behavior by exploiting cognitive vulnerabilities, often without individuals' awareness or against their interests. While "manipulation" is a loaded term, it's crucial to distinguish between acceptable influence—such as persuasion—and problematic forms like *deception or exploitation*. AI's capacity for data-driven decision-making makes it easier to cross into manipulative territory, as seen in the Cambridge Analytica scandal. In public policy, AI could enable **"dark nudges,"** subtly coercing individuals by exploiting psychological triggers or shaping opinions without transparency, thus undermining democratic legitimacy. This blurring of lines between *guiding and controlling* behavior poses risks to citizens' autonomy and informed consent. To mitigate this, the adoption of ethical guidelines, legal safeguards, and oversight mechanisms, such as the EU AI Act's ban on manipulative AI, is crucial. The concept of **"cognitive security"** calls for protecting individuals' decision-making from intrusive AI influence. Rigorous ethical review and public scrutiny are necessary to ensure AI interventions remain aligned with democratic principles.

4. Policy Recommendations & Solutions

4.1. Strategies to Ensure AI Accountability in Policymaking

To harness AI for public good while avoiding pitfalls, governments need robust accountability mechanisms around AI policymaking, holding stakeholders responsible for outcomes, errors, or harms. Several strategies emerge from the literature and policy experiments:

- **Algorithmic Impact Assessments (AIA)**: Prior to AI deployment, agencies should conduct impact assessments, akin to environmental evaluations, to address risks such as bias, privacy violations, and data integrity. Canada's mandatory AIA serves as a model.
- **Auditing and Continuous Monitoring**: Ongoing audits, both technical and procedural, ensure that AI systems perform as intended and adhere to ethical standards. Independent auditors or external agencies enhance credibility and objectivity.

- **Human Oversight ("Human in the Loop")**: Clear accountability lines must ensure human responsibility for AI-driven decisions. Public servants should be equipped to interpret AI outputs and intervene when necessary.
- **Documentation and Record-Keeping**: Comprehensive records, detailing AI system development, objectives, and decisions, support transparency and accountability.
- **Legal and Regulatory Oversight**: External frameworks, such as the EU AI Act, establish mandates for impact assessments and citizen rights to challenge AI decisions, ensuring robust accountability structures.

Investing in expertise and fostering a culture of transparency ensures that AI serves the public good while maintaining accountability and trust.

4.2. Frameworks for Cognitive Security and Guarding Against Manipulative Nudges

To safeguard citizens from manipulative nudging, the emerging concept of **cognitive security** aims to protect mental autonomy and integrity in the face of AI influence. In view of providing a robust structure for its implementation, several initiatives stand out:

- **Ethical Guidelines**: Governments should adopt charters prohibiting manipulative practices, such as *dark nudges*, ensuring AI nudges are transparent, reversible, and beneficial to individuals, not just the state.
- **Restrictions on Micro-Targeting**: Limiting hyper-personalized messages in sensitive domains, ensuring nudges are public and avoid creating divergent realities for different groups.
- **Cognitive Impact Assessments**: Policies should evaluate potential cognitive manipulation, ensuring nudges respect autonomy and offer easy opt-outs.
- **Education and Digital Literacy**: Raising public awareness about AI nudging techniques enhances citizens' resilience to manipulation and promotes critical thinking.
- **Legal Protections for Mental Autonomy**: Legal frameworks could protect *freedom of thought*, challenging intrusive AI influence on decision-making.

Fostering cognitive security and sovereignty requires prioritizing empowerment over coercion. A secure approach enables informed choices, presenting balanced information while avoiding AI-driven distortions. In essence, AI should be a facilitator of autonomy, not a tool of covert control. By formalizing the above-mentioned protective policies, governments can reassure citizens that AI nudges won't become a digital "brainwashing" apparatus but will remain within ethical bounds.

4.3. Citizen Engagement Strategies for AI Transparency and Inclusion

For AI-driven policymaking to be truly democratic, citizens must be active participants in molding and approving its algorithms, not just mere subjects. **Citizen engagement** is thus vital for legitimizing AI in governance, manifesting itself through a plethora of initiatives including:

- **Public Consultations and Co-Design**: Governments should hold forums and invite citizens to co-design AI systems, ensuring public concerns and values shape their implementation.
- **Deliberative Democracy Methods**: Citizens' Juries or Assemblies can deliberate on AI policies, guiding policymakers on acceptable AI applications. Such methods build **legitimacy** and trust in AI.

- **Transparency Portals/Dashboards**: Governments should set up transparency portals, publishing performance metrics and audit summaries for public scrutiny, signaling openness and inviting feedback.
- **Feedback and Redress Channels**: Citizens need clear channels to challenge AI decisions, supported by advisory boards or helpdesks for human review and resolution.
- **Public Awareness Campaigns**: Governments should educate the public on AI's role, benefits, and safeguards, demystifying its use through storytelling and transparency about limitations.

Engagement recognizes that **governance is a social process**, ensuring AI aligns with societal values and builds public trust through collaboration.

5. Conclusion

AI-powered policymaking is at a crossroads—offering groundbreaking possibilities but also raising deep ethical and governance challenges. Governments are increasingly leveraging AI-driven nudges to guide behavior, making public policies more effective and adaptive. However, as AI personalizes interventions and optimizes decision-making, it also tests the boundaries of transparency, autonomy, and public trust.

The European Union's AI Act takes a strict regulatory approach, drawing clear red lines around manipulative AI practices and ensuring high-risk AI systems remain transparent and accountable. Singapore, on the other hand, follows a principles-based model, emphasizing ethical guidance and industry-led best practices. Both frameworks reflect the urgent need for AI governance, but they take different routes—one leaning on enforcement, the other on adaptability.

Public trust in AI-driven governance is fragile but essential. If people feel AI systems serve them fairly and transparently, they will embrace them. If AI is seen as opaque or manipulative, resistance will grow. Ethical concerns—algorithmic manipulation, cognitive security, and fairness—must be tackled head-on. Governments need strong accountability measures, independent oversight, and clear ethical guardrails to ensure AI serves society rather than undermines it.

But AI policymaking shouldn't just be a top-down endeavor—it needs public engagement. involving citizens in discussions about AI governance, implementing transparency measures, and providing clear avenues for redress will strengthen legitimacy. AI should help people make better choices, not make choices for them. The challenge ahead is balancing AI's efficiency with democratic values and human agency—ensuring that as AI reshapes governance, it remains a tool of empowerment, not control.

The future of AI policymaking isn't about resisting AI's potential—it's about using it responsibly, fairly, and transparently. With the right balance of innovation and oversight, AI can enhance governance while safeguarding the freedoms and accountability that define democratic policymaking.

References

[1] Institute for Government. "Nudge Unit." *Institute for Government*, 2024. Available at: https://www.instituteforgovernment.org.uk/article/explainer/nudge-unit.

[2] Sunstein, C. and Thaler, R. "Nudge: Improving Decisions About Health, Wealth, and Happiness." *Yale University Press*, 2008.

[3] Harris, D. "Nudge 2.0: Addressing Medical Diagnostic Error in Heart Disease." *University of Chicago Policy Research*, 2024. Available at: https://harris.uchicago.edu/nudge2.0.

[4] Ali, M. and Khan, Z. "AI-Augmented Nudges: A New Frontier for Tailored Interventions in Sustainability." *ResearchGate*, 2023. Available at: https://www.researchgate.net/publication/AI_Augmented_Nudges.

[5] Yeung, K. "The Start-Ed Programme: AI-Driven Nudging and the Erosion of Individual Autonomy." *Melbourne Law School*, 2023. Available at: https://law.unimelb.edu.au/start-ed.

[6] NCC Group. "Understanding the EU AI Act: A Comprehensive Guide." *NCC Group*, 2024. Available at: https://www.nccgroup.com/eu-ai-act-guide.

[7] European Commission. "EU Artificial Intelligence Act, Article 5: Prohibited AI Practices." *Official Journal of the European Union*, 2024. Available at: https://artificialintelligenceact.eu/article-5.

[8] Personal Data Protection Commission (PDPC). "Singapore's Approach to AI Governance: Model AI Governance Framework." *PDPC*, 2024. Available at: https://www.pdpc.gov.sg/AI-Governance.

[9] OECD. "Algorithmic Transparency Recording Standard." *OECD AI Policy Observatory*, 2024. Available at: https://www.oecd.ai/algorithmic-transparency.

[10] Pew Research Center. "Americans' Views of Artificial Intelligence." *Pew Research Center*, 2024. Available at: https://www.pewresearch.org/fact-tank/ai-public-opinion.

[11] UK Government. "Public Attitudes to Data and AI: Tracker Survey (Wave 4) Report." *GOV.UK*, 2024. Available at: https://www.gov.uk/public-attitudes-data-ai.

[12] Ipsos. "Deliberative Research on AI: Bringing the Public Voice into Decision Making." *Ipsos Policy Research*, 2024. Available at: https://www.ipsos.com/en/deliberative-research-ai.

[13] Ali, T. "Use of Artificial Intelligence to Enable Dark Nudges by Transnational Corporations." *NCBI PMC*, 2024. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC-dark-nudges.

[14] OECD. "Algorithmic Impact Assessment (AIA): A Tool for AI Regulation." *OECD Science, Technology and Innovation Policy*, 2024. Available at: https://stip.oecd.org/aia-assessment.

[15] Government of Canada. "Algorithmic Impact Assessment - Open Government Portal." *Government of Canada*, 2024. Available at: https://open.canada.ca/data/algorithmic-impact-assessment.

[16] White Case LLP. "AI Watch: Global Regulatory Tracker - Singapore." *White Case*, 2024. Available at: https://www.whitecase.com/publications/ai-watch-singapore.