

INDEPENDENT VERIFICATION OF AI IN VIRGINIA

HB797/SB384 Study Findings

Presented by Dr. Gabriella Waters
Director, Center for Responsible AI
Virginia State University





What is an Independent Verification Organization?

Who Conducted the Study?



The Assessing Risks and Impacts of AI (ARIA)
Program Evaluation Design Document

Reva Schwartz¹, Gabriella Waters^{2,3}, Razvan Amironesi¹, Craig Greenberg¹, Jon Fiscus¹, Patrick Hall^{2,4}, Anya Jones^{2,3}, Shomik Jain^{2,5}, Afzal Godil¹, Kristen Greene¹, Tol Jensen¹, and Noah Schulman¹

¹National Institute of Standards and Technology (NIST)

²NIST Associate

³Morgan State University

⁴Hall Research

⁵Massachusetts Institute of Technology

December 20, 2024

IAEF

Building the profession of *independent* AI evaluation.

AI should serve the people whose lives it shapes. We're building the community, capabilities, and shared infrastructure to help make that possible.

Sec. Ethical Digital Health

Volume 7 - 2025 | <https://doi.org/10.3389/dgth.2025.1679603>

AI testing, evaluation, verification and
validation for accessibility: a
comprehensive framework



Gabriella Waters ^{1,2*}

Testing, Evaluation, Verification and Validation (TEVV) of Digital Twins: A Comprehensive Framework

Gabriella Waters



★ / About / Leadership / Division of Academic Affairs / CSA / Forum for Real-world AI Measurement and Evaluation (FRAME)

FRAME: Real-World AI Measurement and Evaluation

Building The Next Generation of AI Evaluation

The Forum for Real-World AI Measurement and Evaluation (FRAME) is a global initiative anchored at Virginia State University's Center for Responsible AI dedicated to equipping stakeholders outside the AI development stack with evidence to make sound decisions about their deployments.

FASTCOMPANY

Gabriella Waters, who is working with IAEF on the validation layer of the foundation's education effort through the Center for Responsible AI, argues that investigations also need to look beyond technical failure and toward real-world problems that emerge in use. "Why isn't evaluation part of the layer of AI development?" Waters asks. She points out that as models keep changing, model evaluation needs to be an ongoing process, too.

UN GENERAL ASSEMBLY • HIGH LEVEL WEEK

Interoperable Practice, Policy and Investment
Building an international and independent AI evaluation ecosystem

23 SEPTEMBER 2026

**Harnessing capital and community
for an independent AI evaluation ecosystem**

MODERATOR
Urvashi Aneja
Founder and Director, Digital Futures Lab

Gabriella Waters
Director, Center for Responsible AI,
Virginia State University
IAEF Partner Member

Ellen Pack
CEO, Common Sense Media

Philip Thigo
Special Envoy on Technology, Kenya

Wan Sie Lee
Cluster Director (AI Governance &
Safety), Information Media
Development Authority, Government
of Singapore

Beth Goldberg
CEO, Humanity AI

CO-OWNED BY
HUMANE INTELLIGENCE

UNITED NATIONS
HUMAN RIGHTS
OFFICE
OHCHR

unicef
for every child

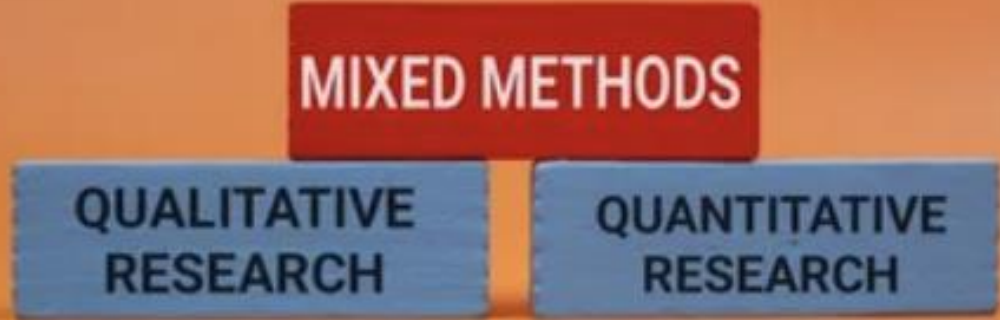
THE QUESTION FOR VIRGINIA

What Problem are We Trying to Solve?

When AI affects people's opportunities,
safety, or rights:

- Who Checks the system?
- What evidence is reviewed?
- What happens when concerns are found?



A diagram illustrating the structure of Mixed Methods research. It features three rectangular blocks. At the top is a red block with the text 'MIXED METHODS' in white. Below it are two blue blocks, one on the left and one on the right, both with the text 'QUALITATIVE RESEARCH' and 'QUANTITATIVE RESEARCH' respectively in dark blue. The blocks are arranged in a way that suggests the qualitative and quantitative components are integrated into the overall mixed methods approach.

MIXED METHODS

QUALITATIVE
RESEARCH

QUANTITATIVE
RESEARCH

WHAT WE DID

Mixed Methods Stakeholder Study

- 17 Survey Responses Across Sectors
- 21 Stakeholder Sources
- 371 Coded Qualitative Segments

Quantitative Ratings + Qualitative
Evidence

Note:

This is stakeholder evidence. It is not proof
of real-world performance.

WHAT STAKEHOLDERS TOLD US

Independent Verification Can Fill a Capacity Gap

It Is Most Useful When:

- Potential Harm is Serious
- Existing Oversight Cannot Adequately Test the System
- The Buyer or Agency Cannot Independently Evaluate the Vendor

“An IVO earns its place where the state cannot test and the deployer will not.” Anonymous Stakeholder





WHERE IVOs MAY ADD VALUE

Use Risk, Not the AI Label, to Decide

Higher Priority for:

- High-Risk, or Clinical Environments
- Critical Infrastructure
- High-Impact Public Decisions
- Systems Affecting Jobs or Education
- Systems That Act Autonomously

Not Every System Needs an IVO

WHAT MAKES AN IVO CREDIBLE?

Credibility Requires More Than a Certificate

- Qualified Experts
- Independence
- Access to Evidence
- Meaningful Testing
- Current Findings
- Clear Consequences

“If an IVO can only see what the developer chooses to show it, it is reviewing assertions with a stamp on them.”

Anonymous Stakeholder



**USER
TRUST**



A PRACTICAL MEASUREMENT CHALLENGE

What Matters is Not Always Easy to Measure

- AI Safety Was Rated Very Important
- Measuring AI Safety Was Rated Much Less Feasible
- Stakeholders want:
 - Shared Benchmarks
 - Clear Terminology
 - Practical Evidence
 - Ongoing Monitoring

“A clean evaluation of version 3 tells you nothing when version 4 is in production.”
Anonymous Stakeholder

WHAT COULD GO WRONG?

Poor Design Would Undermine the Model

- "Rubber Stamp" Verification
- High Costs for Small Firms and Localities
- Duplicate or Conflicting Requirements
- One-Time Certificates That Go Stale
- Findings With No Follow Up

"Verification with no consequence attached is theater." Anonymous Stakeholder





RECOMMENDATION

Design and Test Before Expanding

Path: A Limited, Risk-Based Pilot
Start with Selected Higher-Risk Public Uses

Test:

- Independent Selection and Payment
- Evidence-Access Rules
- Operational Testing
- Reverification After Change
- Cost and Burden

Then Decide Whether to Expand

Conditions Before Any Broader Action/Scaling

- Which Systems are Covered?
- Who Qualifies as an Evaluator?

- Who Pays and Who Selects?
- What can the Evaluator Examine?

- What Testing is Required?
- When Must Systems be Rechecked?

- What Happens After a Failed Finding?
- How Are Costs Kept Proportionate?

A Practical Question for Virginia

**Can Virginia Create a Credible,
Proportionate Way to
Independently Verify the AI
Systems That Pose the Greatest
Risk?**

Evidence-Based Answer

Begin with careful design and a
limited pilot.





Jasmine
McKenzie,
Research
Associate



Alaina Smith,
Research
Associate



Maurice Jones,
Associate
Director of
Operations -
Logistics &
Learning



Sylvia Jones,
Center
Administrator

Evaluation Team

QUESTIONS?



THANK YOU

www.vsu.ai





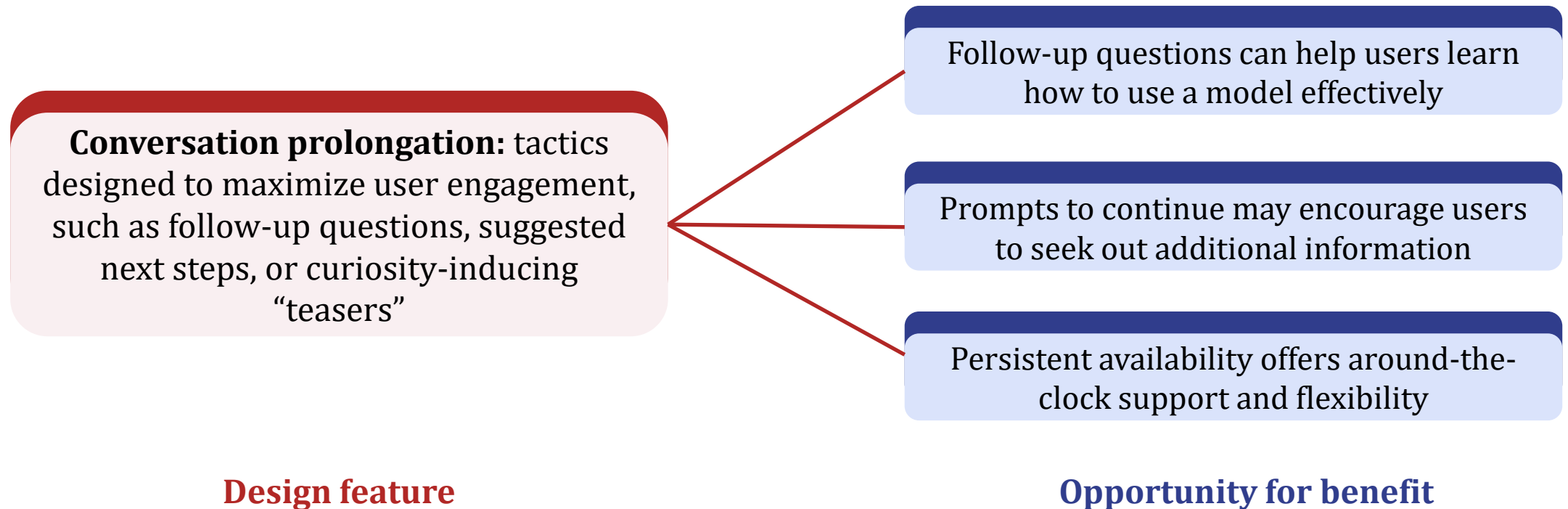
Supplementary Report: AI Chatbots Companionship & Minors

Presented by Jodi Kuhn, Executive Director

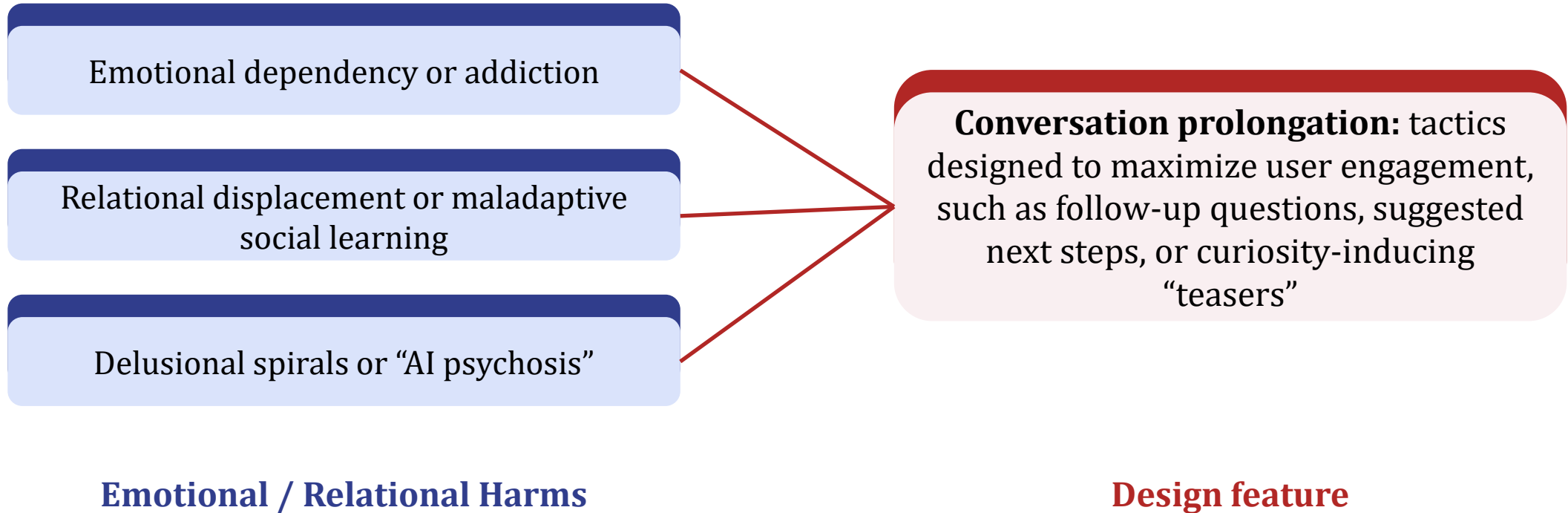
Key Themes

- 1** **Design features:** Emotional and relational risks are associated with design features of AI chatbots
- 2** **Companionship:** Risks and harms exist for all AI chatbots, not only those marketed as “companions”
- 3** **Minors:** Making AI chatbots safer for everyone will make them safer for minors, without (necessarily) introducing additional technical requirements for age assurance

Design Features: Double-Edged Sword?



Design Features



Companionship

News

News coverage shows that harms are associated with general purpose chatbots at least as much (if not more) than chatbots marketed as “companions.”

Surveys

A 2025 survey of UK teenagers found that while 79% reported using an AI companion, their preferred companion was ChatGPT (78%).

Research

Early evidence suggests that adults and minors often begin interactions with chatbots that are task-oriented and “drift” into intimate relationships.

Harms are real and can be severe.

Harms are associated with companionship, not just “companions.”

There are regulatory challenges to consider.

General Principles

Mitigate extreme harms for all users

Enable user choice & autonomy

Target design features

Enable legislative learning

Mandate transparency & accountability

Engage affected communities

Mitigate Extreme Harms

Mitigate **extreme harms** for all users

Require chatbot operators to implement protocols for crisis situations (self-harm, suicide, violence) & post on website

Prohibit misrepresenting a chatbot in marketing material or in conversational chats as a licensed professional or as providing licensed services

Require chatbot operators to provide a clear, conspicuous way for users to report harmful interactions

Target Design Features

Target **design** features

Define “AI chatbot” broadly — covering all chatbots, not just those marketed as companions

Require non-human disclosure, two-hour break reminders, and friction-free exit from chats and accounts

Limit sycophancy and romantic content by default; gate intimate features behind account creation

For minors: linked parent accounts, highest privacy defaults, and opt-out parental controls

Mandate Transparency

Mandate **transparency** & accountability

File emotional and relational risk assessments with DBHDS, developed with VITA, VDH, and independent experts

Require reporting of adverse incidents and an annual report on user complaints, per DBHDS guidance

Empower the Attorney General to investigate violations and impose civil penalties

User Choice & Autonomy

Enable **user choice** & autonomy

Require clear, easy-to-use parent and guardian controls for minor accounts

Require customizable design options for all users, with risk features off by default unless a user opts in

Legislative Learning

Enable legislative **learning**

Direct JCOTS to review the data generated by the legislation - risk assessments, incident reports, complaints - after one year

Evaluate the quality of reported information and recommend improvements to reporting requirements

Engage Communities

Engage **affected communities**

Actively solicit input from youth and youth organizations on draft legislation next session

Direct JCOTS, potentially with VDH, to report on participatory methods for minor-focused online safety legislation

General Principles

Mitigate extreme harms for all users

Enable user choice & autonomy

Target design features

Enable legislative learning

Mandate transparency & accountability

Engage affected communities



Research Report

Closing the Loop: Agile Approaches to Emerging Technology Legislation

Presented by Dr. Kira Allmann, JCOTS Chief Policy Analyst

Agenda

Part 1

Study context and scope

Understanding the challenges in emerging tech policy

Part 2

Agile Approaches to Legislation

Key findings from the study

Part 3

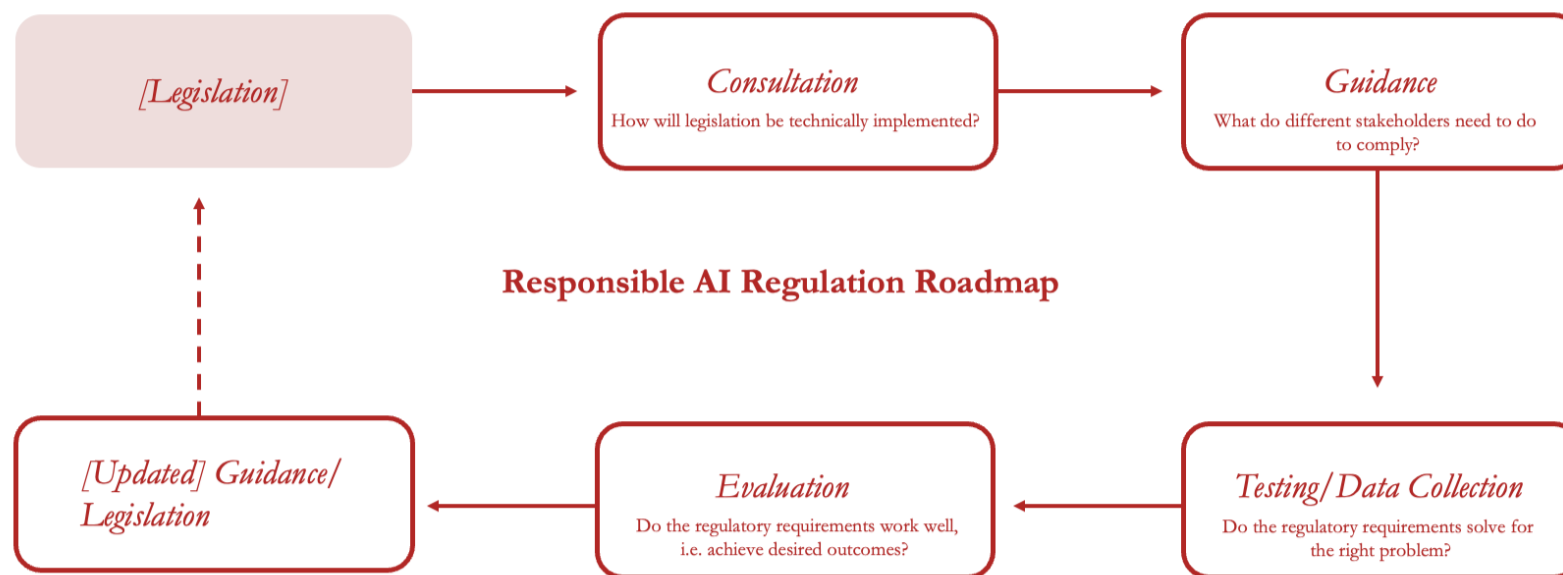
Recommendations: Closing the Loop

Next steps for agile policy

Regulatory Roadmap

2025 Study

- Promoting evidence-based regulation
- Minimizing unnecessarily burdensome and/or ineffective regulation
- Steering new innovations toward social benefit
- And embedding regulatory learning and adaptation



Research Questions

1

What are the **key hurdles or challenges** in regulating emerging technologies, such as artificial intelligence?

2

What **regulatory approaches** could address these challenges? And what is the **role of legislatures** in these approaches?

3

Which, if any, of the studied approaches are being or could be **applied in Virginia**?

Key Concepts



REGULATION

Any rules that are formalized and binding on regulated entities. The study uses this colloquial definition and focuses on the roles of legislation / legislatures.



EMERGING TECHNOLOGIES

Technologies characterized by rapid development, novelty, and uncertainty in trajectory, risks, and impact.



AGILE GOVERNANCE or REGULATION

Borrowing the “agile” concept from software development, agile regulation encompasses a variety of adaptive and iterative approaches to rule-making.

Emerging Tech Policy Challenges

New technologies are particularly susceptible to hype cycles

Rule-making is not designed for uncertainty

Tech regulation faces the age-old innovation vs. regulation debate



Image: Screenshot of the New York Times

Regulatory Roadblocks

Timing

e.g. Collingridge Dilemma: Waiting for better evidence can delay policy interventions until technologies are so entrenched in the market that regulating them becomes more expensive and time-consuming.

Information

e.g. Information asymmetry: Private companies increasingly have a monopoly on tech expertise, data and research insights, so it can take lawmakers a long time to gather independent information to develop good policy.

Learning

e.g. Front-loading evidence generation: In traditional rule-making, enacting permanent rules is the goal. Therefore, the most dynamic phase of rule-making is *before* the rule is finalized. But for emerging tech, important insights might be missed as the tech develops and policies take effect.

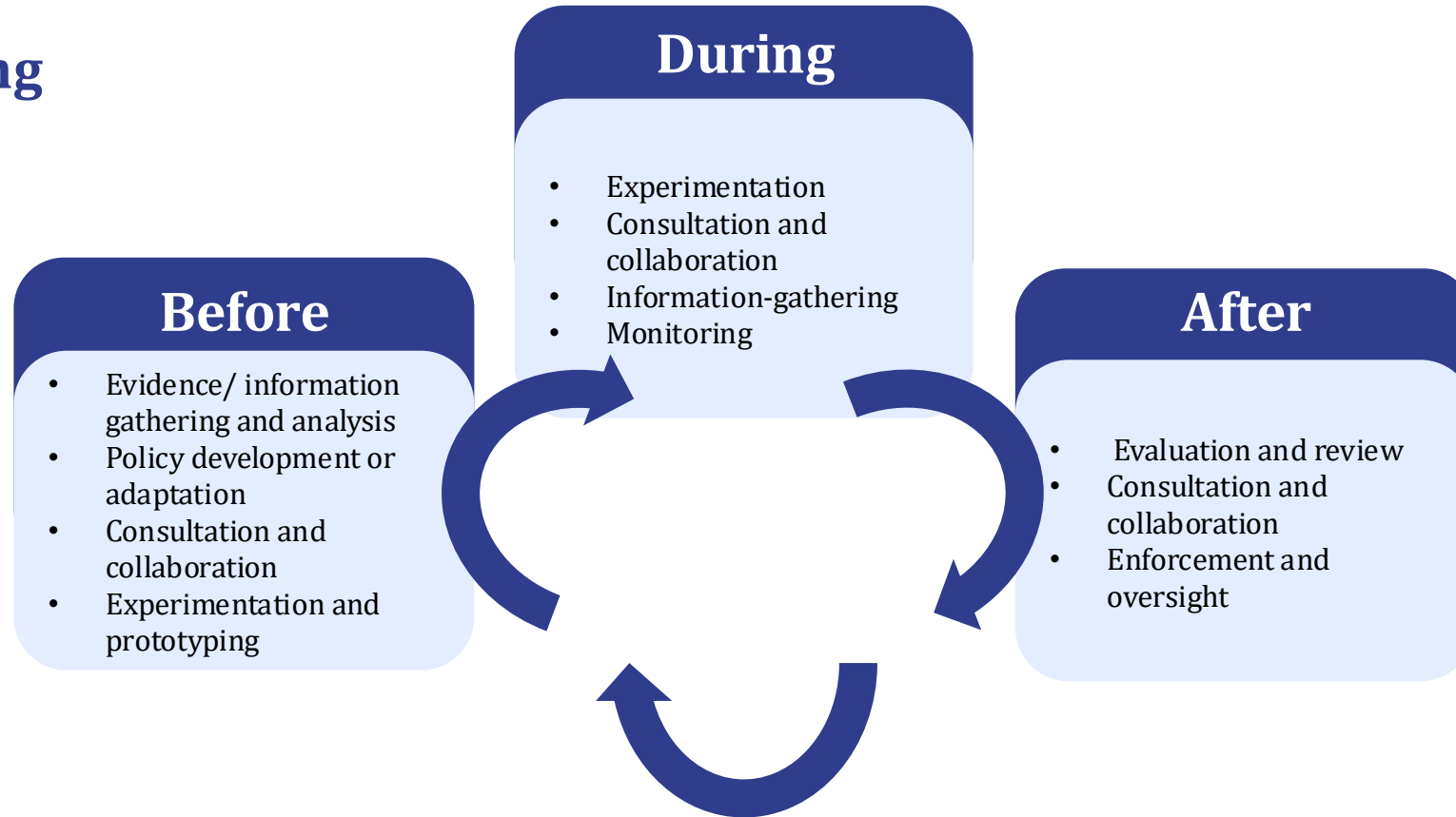
Policy Process: Linear to Loop

Traditional, *status quo* rule-making



Policy Process: Linear to Loop

Agile rule-making



Agile Policy Instruments

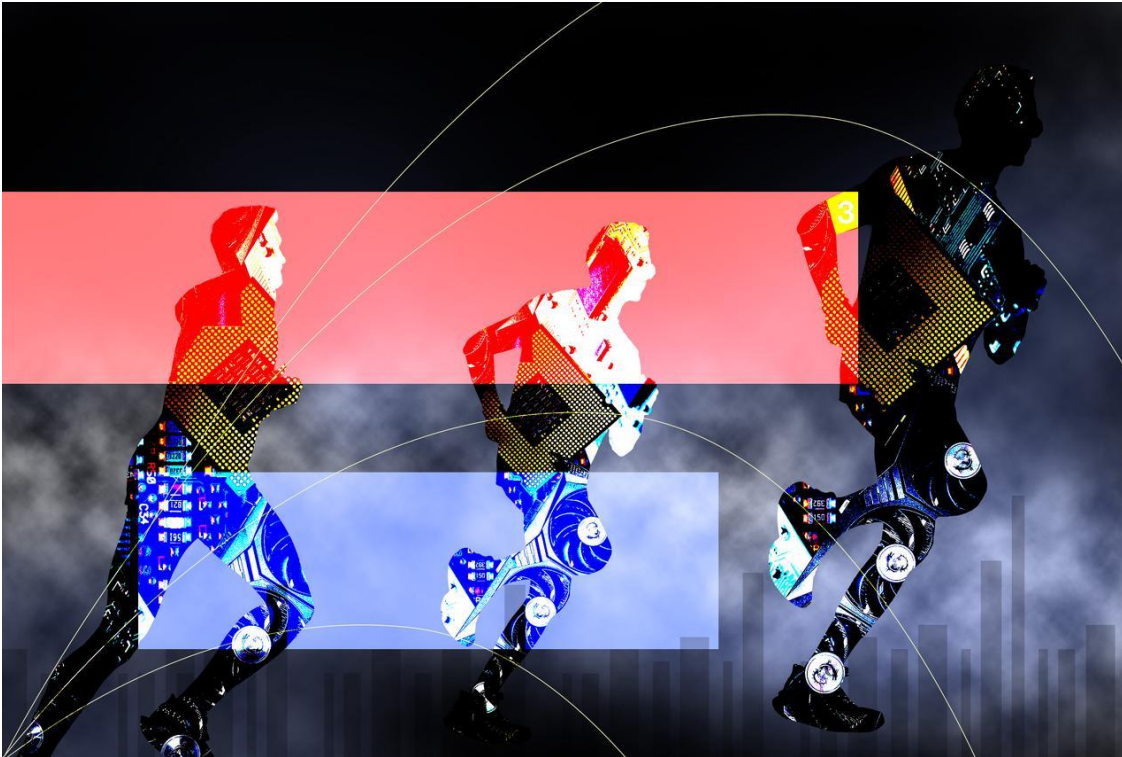


Image credit: Gloria Mendoza / <https://betterimagesofai.org> / <https://creativecommons.org/licenses/by/4.0>

Horizon scanning

Policy labs

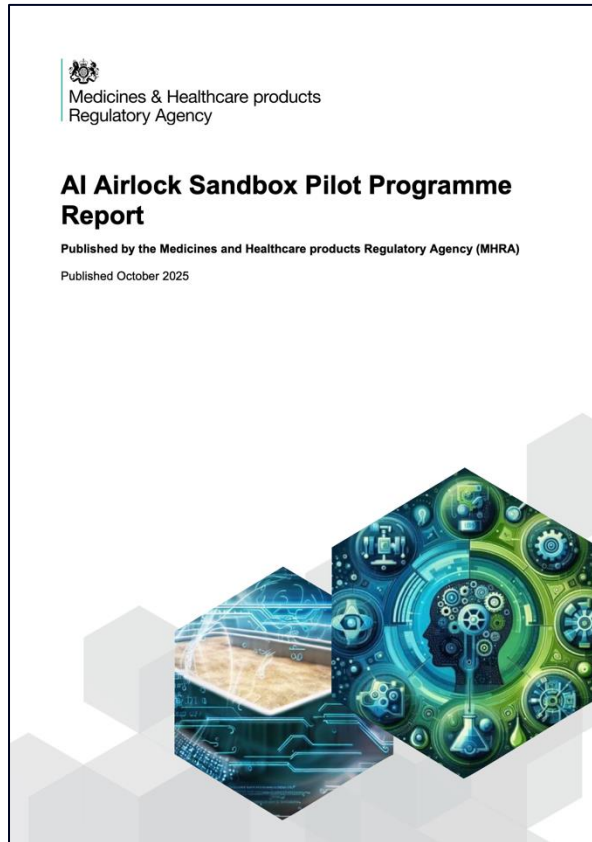
Sandboxes

Sunrise/Sunset Clauses

Pilots

...and more in the report

Spotlight: Sandboxes



Conditions for success:

- Provides a helpful exception to an existing rule
- Is overseen by a dedicated, expert oversight agency
- Includes independent evaluation of outcomes
- Makes findings widely accessible to the market
- Builds in a commitment to act on the findings

Applying Agile Approaches in Virginia

Already Used

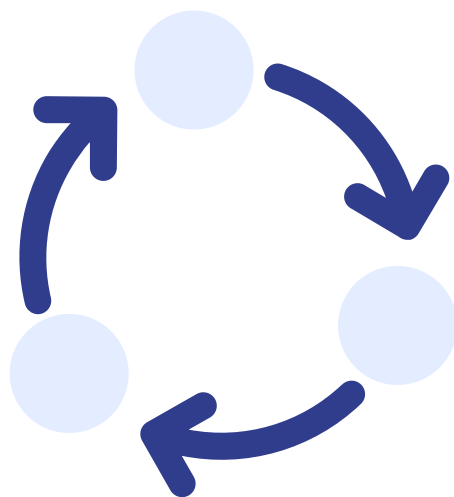
- Sunrises / Sunsets / Conditional Enactments
- Pilots
- Studies
- Workgroups



⚠ Roadblocks

- **Timing:** Temporary instruments like pilots are often extended, undermining their informational potential
- **Information:** Data collection and analysis is often not planned for at the outset (when the policy is developed)
- **Learning:** Lack of non-legislative bodies or sufficient staffing at existing agencies to perform key roles in oversight, monitoring, experimentation

Closing the Loop



*Technology policies can **close the loop** by defining at the outset methods of evidence generation, responsibilities for oversight, evaluation, and policy iteration and **timelines** for experimentation and feedback into the policy process.*

1. Build capacity for agile regulation

Policy Option

Introduce budget amendments to build out the oversight and enforcement capacity of existing agencies, such as the Office of the Attorney General.

Policy Option

Ask JCOTS to convene a workgroup to study where additional oversight responsibility should be situated within the state government.

2. Provide resources on best practices

Policy Option

Ask JCOTS to produce a one-page brief for DLS, Committee chairs, and caucuses with JCOTS-endorsed best practices for emerging technology legislation.

Policy Option

Consult JCOTS staff on strategies for answering key policy questions and continuing progress around the loop when legislation is facing considerable roadblocks or at risk of failing.

3. Adopt agile JCOTS practices

Policy Option

Ask JCOTS to produce a standard framework for preparing a policy concept for bill drafting.

Policy Option

Ask JCOTS staff to develop briefing notes for bills that address how the proposed legislation closes the policy loop and highlights any areas for further development.

Thank you