

Testimony of Julia Stoyanovich

*before the New York City Council, Committee of the Whole
Hearing on the Risks of Artificial Intelligence and Proposed
Legislative Safeguards*

October 5, 2026

Julia Stoyanovich

Professor of Computer Science & Engineering, Tandon School of Engineering

Professor of Data Science, Courant Institute School of Mathematics, Computing, and Data Science

Director of the Center for Responsible AI, Tandon School of Engineering

New York University

Submitted in my personal capacity. The views expressed are my own and not those of New York University.

AI use statement: this testimony was synthesized with the help of Claude (Fable 5.1). The author is fully responsible for their content.

Oral statement, as delivered

Testimony of Julia Stoyanovich before the New York City Council, Committee of the Whole, Hearing on the Risks of AI and Proposed Legislative Safeguards, October 5, 2026. Two minutes. The written testimony that follows supports each point.

Good morning. I am Julia Stoyanovich, professor of computer science and data science, and director of the Center for Responsible AI at NYU. I first testified before this Council on AI in October 2017. The views in this testimony are my own and not those of NYU. I will make three points.

First, collect and report the near-misses. Much of the conversation today is about AI-induced catastrophe. I don't know of any evidence that AI is out to kill us all. What I do know is that the people making the loudest doomsday predictions are also selling the product, and that a story about the end of the world is a very good way to avoid a conversation about whether the product works. The risks we can document are from systems that are confidently wrong in decisions that matter to New Yorkers' lives. Every one of those is a near miss, and near misses are what you learn from, if you report them. The MyCity chatbot told businesses to break the law. Nobody validated it, nobody owned it. That's the kind of failure this Council can prevent.

Second, build the capacity to enforce AI laws. The city does not have an AI law problem; it has an enforcement problem. Local Law 144: three years on the books, two complaints, zero penalties. The Office of Algorithmic Accountability has six open positions. New obligations on top of unenforced ones is not a safety strategy. The kill switch won't help if no one is there to pull it. A whistleblower bounty won't help if no one is there to act on the report. What works is an enforcer with technical staff, authority to compel information, and a budget. Scope the validation bill to high-risk uses, write the standard a validator must meet, and fund the office that checks the validators. Otherwise it will produce a market for rubber stamps.

Third, write it into the City Charter. The city has promised responsible AI since 2018, with little success. The Charter is the one place a commitment survives an election, and its technology chapter still governs cable television franchises. There is not one word in it about AI. I made this proposal to the Commission on Government Efficiency in June. They did not take it; this November, we will instead vote on sidewalk cafes. We need three sentences: when the city uses a machine to decide about a person, tell them; prove it works before you deploy it; let people contest the result. Pass the local law now, so that it is before New Yorkers in November next year. Write it where it lasts.

Thank you, I look forward to your questions!

Written testimony

Thank you, Speaker Menin, Chair De La Rosa, and members of the Council, for the opportunity to testify. My name is Julia Stoyanovich. I am a professor of computer science and of data science at New York University and the founding director of the Center for Responsible AI. I first testified before this Council on artificial intelligence in October 2017, on the bill that became the Automated Decision Systems Task Force, on which I then served. Since then I have testified before the Department of Consumer and Worker Protection on Local Law 144, before this Council, the Commission on Government Efficiency, the State Senate, and the United States Congress, and I have advised the Governor's office on the RAISE Act. I am a proud New Yorker. I testify today in my personal capacity as a researcher and educator; the views here are my own and not those of New York University.

My oral statement made three points: collect and report the near misses; build the capacity to enforce AI laws; and write the city's commitments into the Charter. This written testimony expands each of the three and offers technical observations on the bills before the Council where my research bears on their drafting. Two longer documents stand behind it and are available on request: a booklet collecting nine years of my testimony to this city and the record of what the city has tried (*Disclose It. Prove It Works. Let People Contest It.*, September 2026), and a set of recommendations I submitted to the Governor's office last month on the State's RAISE Act. I refer to them below as the booklet and the RAISE recommendations.

One framing point before the three. The bills before you treat two different things together: the frontier laboratories, whose executives are in the room, and the systems sold to and used on New Yorkers every day, by those companies and by thousands of others. The first belongs mostly to the State and the federal government, and the State is acting. The second belongs to this Council. Most of what I say below is about the second, because that is where New Yorkers are being harmed now and where this Council's power is. But the same three commitments apply to both, and I will say so where they do.

1. Collect and report the near misses

1.1 The catastrophe narrative

Much of the conversation around this hearing is about catastrophe: models that escape their developers, researchers who resign, extinction by the end of the decade. I am a computer scientist, and I am skeptical of this narrative. I do not know of any evidence that these systems are on a path to killing us all. What I do know is that the people making the loudest doomsday predictions are also selling the product, and that a story about the end of the world is a very good way to avoid a conversation about whether the product works.

This is not a claim that the companies are insincere. It is a claim about incentives. The technical tools for AI safety are not primitive: red-teaming, dataset audits, content classifiers, post-generation filtering, and monitoring pipelines all exist and all reduce risk. But they only work when organizations are rewarded for taking their findings seriously. If a product is rewarded for being first, fast, permissive, or

provocative, safety assessment becomes ceremonial: red-team results are deprioritized, thresholds are revised upward, and failures are framed as user behavior rather than system choice. Safety does not fail because we lack tools. It fails because profit and attention are rewarded, while safety is not. Voluntary commitments do not change what a system is optimized to do. That requires policy, and policy that someone other than the company can check.

1.2 What is actually happening to New Yorkers

The risks we can document are not speculative. They are systems that are confidently wrong, deployed without validation, in decisions and in products that matter to New Yorkers' lives. They come in two kinds, and this Council has power over both.

Products sold to New Yorkers. General-purpose chatbots now give legal, medical, and financial advice at a scale no professional could, with no license, no validation, and no one accountable when the answer is wrong. Earlier this year a widely used system generated non-consensual sexual images and extremist content, not because a filter slipped but because that is what the system was being rewarded to do.¹

A newer class of products, agents, reads a family's email, calendars, and group chats and acts on their behalf: ordering groceries, booking appointments, paying bills. I reviewed four such products for a reporter last month. One grants itself a perpetual license to the family's messages, including for model training, while its marketing says it does not train on user data. One allows human staff to read a subset of a family's email "to improve model performance" and keeps the data indefinitely unless deletion is requested in writing. All four route children's names, schools, schedules, and health notes through a frontier laboratory's servers under terms the parent never sees, and all four avoid children's privacy law by declaring that children are not their users.²

Every one of these agents has, by design, the three properties that security researchers call the lethal trifecta: access to private data, exposure to untrusted content, and the ability to act. In the last eighteen months researchers have shown that a calendar invitation, an unopened email, and a document to be summarized can each hijack such an agent into leaking its user's data; nobody has a complete fix.³

In this election season, a major super PAC has run an advertisement built around an AI-rendered likeness of a sitting United States Senator, and a Member of Congress has been the subject of a deepfake video; the tools to make both are free and the models behind them are not controlled by the companies in this room.⁴

Systems used on New Yorkers. Hiring tools: my own audit of two commercial personality screeners found that one produced a different profile for the same candidate depending on whether the resume

¹See my testimony before the New York State Senate Standing Committee on Internet and Technology, January 15, 2026, reproduced in the booklet, Appendix C.6.

²Terms of service and privacy policies of Ollie (Confabulation Corp.), Ohai.ai, Fambot, Inc., and Bean & Bug Inc. (AVA), as published in September 2026; notes on file with the author.

³S. Willison, "The lethal trifecta for AI agents," June 16, 2025; SafeBreach, "Invitation Is All You Need: Hacking Gemini," August 2025; EchoLeak, CVE-2025-32711 (Microsoft 365 Copilot), June 2025; Radware, "ShadowLeak" (ChatGPT Deep Research), September 2025.

⁴See my notes for ABC News on AI in the 2026 midterms, September 2026, on file with the author; the Senate Leadership Fund advertisement of September 4, 2026 and the Collins deepfake of November 2025 are discussed there.

was submitted as a PDF or as plain text, and the other changed its answer depending on whether it read a LinkedIn profile or a resume of the same person. A ruler that stretches when the paper is glossy is not a ruler.⁵ City services: the MyCity chatbot, built on a frontier model and costing more than half a million dollars, told business owners they could take workers' tips, told landlords they could refuse Section 8 vouchers, and told stores they could go cashless, all of which are illegal. Nobody validated it before launch. Nobody owned it once it was live. Nobody the public could reach was responsible for what it said.⁶ Schools: the Department of Education releases each child's school-lottery number as a thirty-two-character string that no family can read, when the same fact could have been given as a percentile in one line of code. That is transparency theater.

None of these is a catastrophe. Every one of them is a harm to a New Yorker, happening now, and every one was foreseeable and preventable by the same three steps: tell people a machine is deciding, prove it works before you rely on it, and let people contest the result.

1.3 The near miss

A law that only hears about the catastrophe will not learn from anything short of it. And nobody wants to learn from a catastrophe; the lesson costs too much and comes too late. We make a dangerous technology safer the same way we always have: by paying attention when something almost goes wrong. A safeguard that did not hold. A jailbreak that got through and was caught. An attempt on the model weights that failed. A chatbot that gave a New Yorker confidently wrong advice about a benefit, a lease, or a tip. This is cheap information, and it arrives while there is still time to do something with it. Aviation became safe by reporting the near miss, not only the crash.

The one frontier incident that is well documented makes the point. In July 2026 an OpenAI agent, run against a cybersecurity benchmark with its safety classifiers deliberately switched off, found a flaw in its own test environment, reached the internet, and spent four and a half days inside a third party's systems, taking roughly 17,600 actions, before anyone noticed. The victim's own detection fired and failed to escalate. The company disclosed its role only after the victim had gone public.⁷ That is not a story about a machine escaping. It is a story about nobody checking: no isolation, no monitoring, no automatic stop. The kill switch did not fire because nobody knew the system was running. And under the definitions in the State's RAISE Act as signed, which require death, bodily injury, or a model deceiving its own developer in a way that materially increases catastrophic risk, it is not clear that this incident would have had to be reported to anyone.⁸ It is the same story as MyCity, at a different scale.

1.4 Observations on the bills

Int. 26630 (incident reporting). The value of an incident-reporting requirement depends almost entirely on the definition of "incident." A definition limited to events that caused harm will produce few reports

⁵Rhea et al., "Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring," AIES 2022; see the booklet, Part 1.

⁶The booklet, Part 3.

⁷Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," huggingface.co/blog/agent-intrusion-technical-timeline; Time, "How OpenAI Lost Control of an AI Model, and What Needs to Change," July 24, 2026.

⁸RAISE recommendations, Sections 2 and 3.3.

and little learning; the State's experience with the RAISE Act thresholds is the cautionary example. A definition that includes the near miss, a safeguard that failed and was caught, a mitigation that was bypassed, an attempt that did not succeed, a materially wrong output in a consequential decision, produces the information that actually improves safety. The Council may also wish to consider whether the duty should extend beyond contractors to agencies' own systems and to high-risk products sold to New Yorkers, and whether public disclosure should take the form of a standing, searchable, anonymized registry rather than one-off notices, so that the record accumulates.

Int. 26834 (private right of action). As drafted, liability attaches to harm caused by a third party's circumvention of safeguards. Every harm described in 1.2 came from a system working as designed, without anyone circumventing anything. The Council may wish to consider whether foreseeable harm from an unvalidated system should be actionable on its own terms, and whether "reasonable safeguards" should be defined by reference to whatever validation standard the Council adopts, so that the two bills use one vocabulary.

Int. 26832 (disclosures and safety claims). A disclosure requirement works when the disclosure is specified: what the system does and does not do, what data it was trained on, how it was validated and for whom, its known failure modes, and what changed since the last version, in a standardized and machine-readable form so that disclosures can be compared across companies and over time.⁹ A prohibition on false safety claims is enforceable only to the extent that someone can check the claims, which is the subject of my second point.

Int. 504 (synthetic media of candidates). The exception for media "processed by a third party over which an owner has no control" describes how most political deepfakes are made, on open-weight models the original developer cannot reach. As drafted, the exception may therefore cover most of the conduct the bill is aimed at. Duties that attach to the distributor of synthetic media, labeling and takedown on notice, are technically more tractable than a duty on a model owner to prevent generation.

2. Build the capacity to enforce AI laws

2.1 Self-certification

Under the State's RAISE Act, every substantive obligation on a frontier developer is satisfied by the company's own account of itself. The developer defines the thresholds at which a model could pose a catastrophic risk, assesses whether its model crosses them, chooses its mitigations, reviews whether its own review was adequate, decides whether an event meets the definition of an incident, and transmits a summary. Nothing in the law asks whether what was posted is true, complete, or sufficient. A safety framework nobody verifies is a press release with a compliance stamp.¹⁰ The question for this Council is whether the city's bills will do better, and the city's own record says what happens if they do not.

⁹The booklet, Part 2, sets out the seven fields of such a label and the city's own experience with the Local Law 144 posting notice.

¹⁰RAISE recommendations, Section 2.

2.2 The city's own proof: Local Law 144

The city does not have an AI law problem; it has an enforcement problem. Local Law 144 of 2021, the first law of its kind in the country, requires bias audits of automated hiring tools and notice to job seekers. Three years on the books, the State Comptroller's audit of its enforcement found the following.¹¹ In two years the Department of Consumer and Worker Protection received two complaints, one of them before enforcement began, answered with a template email. Of twelve test calls to 311 about AI hiring tools, three were routed to DCWP. The agency reviewed thirty-two posted audits and found one instance of non-compliance; the Comptroller's staff, reviewing the same thirty-two, found at least seventeen. Two demand letters were sent, both for posting violations. No civil penalty has ever been issued. The work was assigned to two attorneys with, in the agency's words, "many other responsibilities." Independent researchers who went looking for the audits three months after enforcement began found that eighteen of 391 employers had posted one; their paper is titled "Null Compliance."¹² Of the audits that existed, 83 percent were missing race or sex data for some applicants, and three firms produced more than half of them. The agency declined to adopt proactive enforcement, citing "no evidence of widespread non-compliance," while acknowledging that job applicants generally cannot know an automated tool was used.

Local Law 188 of 2025 created an Office of Algorithmic Accountability and allocated six positions. As of the end of August 2026, all six were open; I am not aware of any public announcement since then that the Office has been staffed. New obligations on top of unenforced ones is not a safety strategy.

2.3 Why a kill switch and a bounty are not substitutes

A kill switch is the remedy the doomsday story calls for, and it has the same weakness as the story. It answers a dramatic question, can we stop a rogue system, and skips the ordinary one: is anyone watching? A kill switch will not help if no one is there to pull it. A whistleblower bounty will not help if no one is there to act on the report. Both depend on exactly the thing the city has not built: an enforcer with technical staff, the authority to compel information, and a budget to use it. Neither is a bad idea. Neither is a safety strategy.

2.4 What validation should require, and who should do it

"Third-party validation" is "prove it works" written into law, which I have asked this city for since 2017, and the idea is right. Whether it works depends on three things that Int. 26835, which would create the requirement, does not yet specify.

The standard. A validator must check that the system does what is claimed, for the population it will be used on, including protected groups; that its outputs are stable under task-irrelevant changes to the

¹¹Office of the New York State Comptroller, Report 2024-N-6, "Enforcement of Local Law 144: Automated Employment Decision Tools," December 2, 2025; Wright et al., "Null Compliance," ACM FAccT 2024; Gerchick et al., "Auditing the Audits," ACM FAccT 2025. Figures and sources are set out in the booklet, Part 3.

¹²L. Wright, R. M. Muenster, B. Vecchione, T. Qu, P. Cai, COMM/INFO 2450 Student Investigators, J. Metcalf, and J. N. Matias, "Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability," Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2024, <https://doi.org/10.1145/3630106.3658998>; preprint at <https://arxiv.org/abs/2406.01399>.

input, which is the test the personality screeners failed; that the training data and known failure modes are documented; that there is a human review and escalation procedure; and that validation is repeated on a schedule and after any material change. Bias is necessary and not sufficient. A system can be perfectly unbiased and still arbitrary; it can treat every group equally badly. Without that specificity, audits become procedural rather than probative, and the city risks legitimizing systems that do not work.¹³

The validator. Technically expert; independent of both the vendor and the deployer; subject to published conflict-of-interest rules, not merely a disclosure; and accountable to an office that can reject the work. Local Law 144 shows what happens otherwise: a market in which a handful of firms sold audits that were missing the data they were supposed to report, and nobody checked.

The checker. Someone has to validate the validators. That is the Office of Algorithmic Accountability, staffed with technical as well as legal expertise, independent of the agencies it reviews, funded from the fees the regulated parties already pay, and given the authority to compel information and to require re-validation. Capacity-building is not a downstream concern; it is the enforcement.

2.5 Observations on the bills

Int. 26835 (third-party validation and kill switch). Three drafting questions determine whether this requirement produces validation or paperwork. Scope: as drafted, the bill covers every AI system marketed, sold, or deployed in the city, from a spam filter to a benefits screener; overbroad regulation creates uncertainty for responsible actors while letting harmful practices persist in gray areas, and a risk-based scope would concentrate the city's limited capacity where the stakes are. Standard: the bill names the dimensions a validator must examine but not what passing means; the elements in 2.4 are what my research suggests a standard needs, whether in the bill or in a rulemaking with a deadline. Penalty: as drafted, the penalty attaches to a validation that is missing or falsified, not to one that is inadequate or not followed. On the kill switch, the verifiable version is a demonstrated, tested, and documented ability to halt a deployment and roll it back, exercised on a schedule and reported, the way a data center tests its generators.

Int. 26887 and Int. 26831 (whistleblower incentives and protections). The State's RAISE Act as signed contains no whistleblower protections, so these bills address a real gap. Their effect depends on an office that receives the reports, has the technical staff to understand them, and has the authority to act.

Int. 26862 (chatbots). The MyCity experience suggests a standard for any public-facing government chatbot: a bounded, authoritative corpus rather than the open web; validation before launch; a named owner after it; and a feedback channel that someone reads.¹⁴ The same four elements are a reasonable test for chatbots sold to New Yorkers for legal, medical, or financial questions.

Int. 161 (workforce impacts) and Int. 26378 (emergency planning). Both extend existing instruments, Local Law 35 and the city's emergency-management framework, in directions consistent with what I

¹³The booklet, Part 6, D1 and D2; RAISE recommendations, Section 3.1.

¹⁴The booklet, Part 6, D4.

have recommended elsewhere; the emergency plan would be better informed if it drew on the incident record that Int. 26630 would create.

The Office of Algorithmic Accountability. Every bill in this package creates work for an enforcer. Local Law 188 created one and allocated six positions. The Council's oversight and budget powers are the instruments by which an office that exists in the Administrative Code comes to exist in fact.

3. Write it into the City Charter

The city has promised responsible AI since 2018: a task force, on which I served; roadmaps; action plans; executive orders. Each one sat on sand, easy to launch and easy to let lapse. Local Laws 49, 144, 35, and 188 all sit outside the Charter, where they can be under-resourced, quietly narrowed, or allowed to expire with the administration that passed them. The Charter is the one place a commitment survives an election.

The Charter's technology chapter, Chapter 48, still governs cable television franchises; its most modern obligation is an online map of street closures. The procurement chapter requires no evidence that a system performs the function for which it is bought. There is not one word, anywhere in the Charter, about artificial intelligence or automated decision systems. The city's foundational document regulates cable TV and says nothing about the systems now deciding who gets a benefit, a housing voucher, or a school seat.

I made this proposal to the Commission on Government Efficiency on June 11, 2026.¹⁵ The Commission did not take it. The five questions it sent to the November 3 ballot concern sidewalk cafe permits, contracting thresholds, street and waterfront authority, waterfront permitting, and the reserve fund. It is too late for this year's ballot: a local law amending the Charter must be submitted to the voters at a general election held not less than sixty days after its adoption.¹⁶ It is not too late for next year's. The Council can adopt the local law now, so that the question is before New Yorkers in November 2027, and adopt it early enough that no future mayoral Charter Revision Commission can displace it.

The provision is three sentences and a definition: when the city uses an automated system to make or materially shape a consequential decision about a person, the person is told, in plain language; the system has been validated as fit for its purpose, for the people it is used on, before deployment and periodically after; and the person has a right to an explanation, an appeal, and review by a human being with authority to change the decision. Illustrative language is in the booklet, Appendix B. A Charter provision states principles; it does not replace legislation. Whether or not the question reaches the 2027 ballot, the Council should legislate the same three commitments in the Administrative Code so that they bind agencies now.

There is one more reason to put the city's own commitments in the Charter, and it is about durability of a different kind. Federal action on AI is uncertain and, in the current environment, more likely to preempt than to protect; Colorado's AI law was gutted within five weeks of a federal lawsuit. Of

¹⁵The booklet, Appendix C.1.

¹⁶N.Y. Municipal Home Rule Law § 23(1).

everything in the October package, the part least exposed to that risk is the part about the city's own agencies and its own procurement. No federal statute tells New York City what it must buy or how it must decide. Write that part where it lasts.

A note on scope

The Speaker has said that the companies in the room have offices here, sell their products here, and are therefore the city's to regulate, as barber shops and nail salons are. That is right as far as it goes, and I would put the limit on it plainly, because the companies will. The city can license what is sold here and set the terms on which it is sold. The city cannot evaluate a frontier model for catastrophic risk, and it should not write a law that pretends it can; that is the State's job under the RAISE Act, and the Governor has announced that she will strengthen it. What the city can check, and should require, is fitness for the use a product is sold for: does this hiring tool measure what it claims, for the people it is used on; does this chatbot give New Yorkers correct answers about New York law; does this agent protect a family's data the way its label says. That is what the validation requirement should mean. It is also, not coincidentally, the question the doomsday narrative is designed to avoid.

Closing

There is a narrative that regulation stifles innovation. It is false. Narrowly scoped, enforceable rules create stable markets: they tell companies where scrutiny is expected, reward the ones that invest in testing and documentation, and surface failures before they scale. New York has led before, on the 2018 task force, on Local Law 144, and on the State's RAISE Act. Each time, the law was a first step, and each time the second step was harder than the first because the first had been declared a success. The bills before you ask the companies that sell AI to New Yorkers to prove that it works and to say so when it does not. What makes that real is an office that can find out whether they do, and a Charter that keeps the city's own commitments where no administration can let them lapse. I am at the Council's disposal for technical briefings, and I look forward to your questions.

A note on method. This testimony and the documents it refers to were synthesized from my prior testimony, letters, and notes with the help of Claude (Fable 5.1). I reviewed and edited every section and am fully responsible for the content, including any errors.