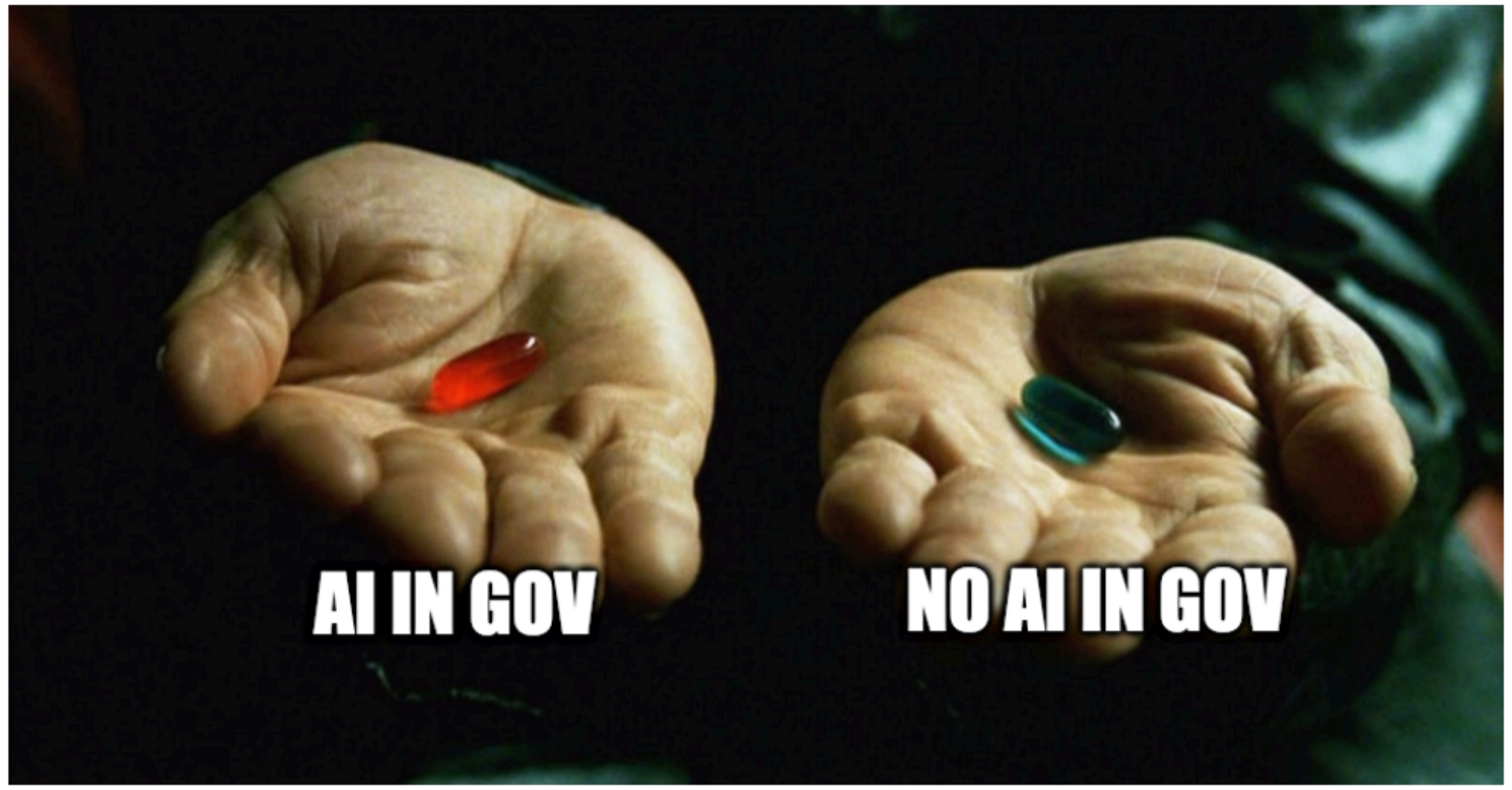
Safety Considerations and Broader Implications for Governmental Uses of AI

Peter Henderson

JD Candidate, Stanford Law School
PhD Candidate, Stanford Computer Science

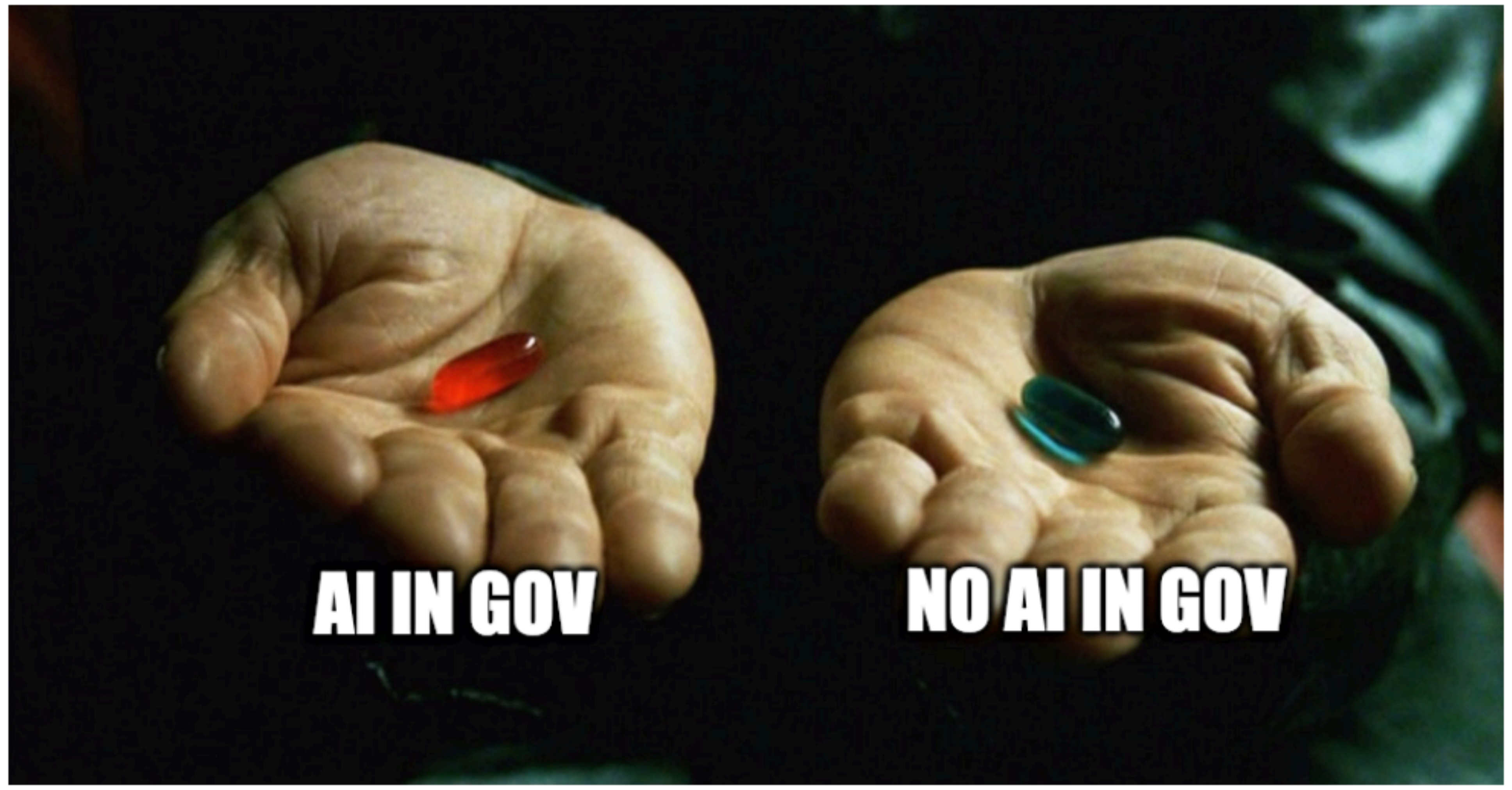
*All views are my own and not of any government agency, company, or other entity.

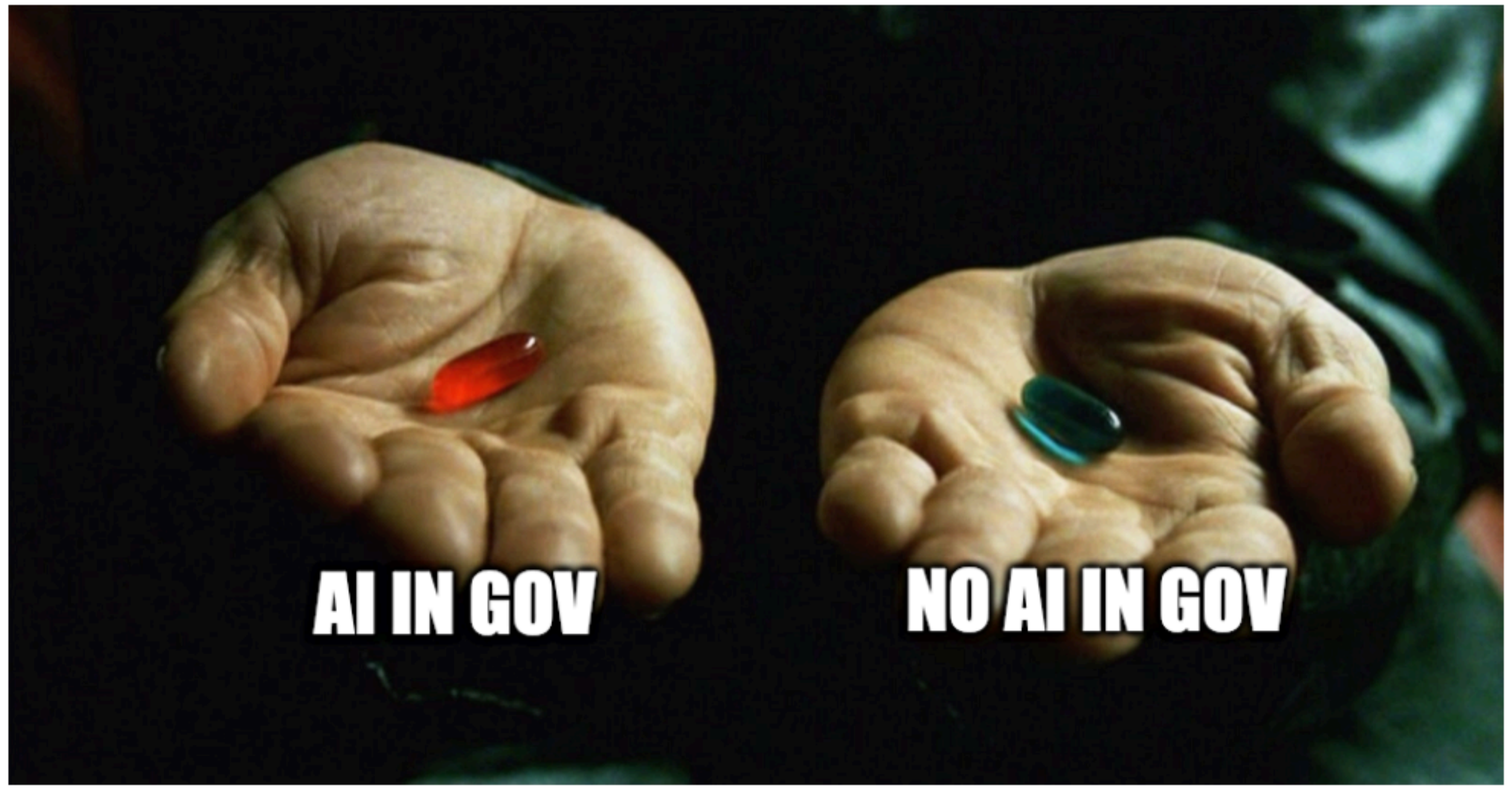

AI IN GOV
NO AI IN GOV

Put AI everywhere as fast as possible!

Humans are terrible at their jobs anyways!

Who needs safeguards?





AI is going to destroy us all, just don’t do it.

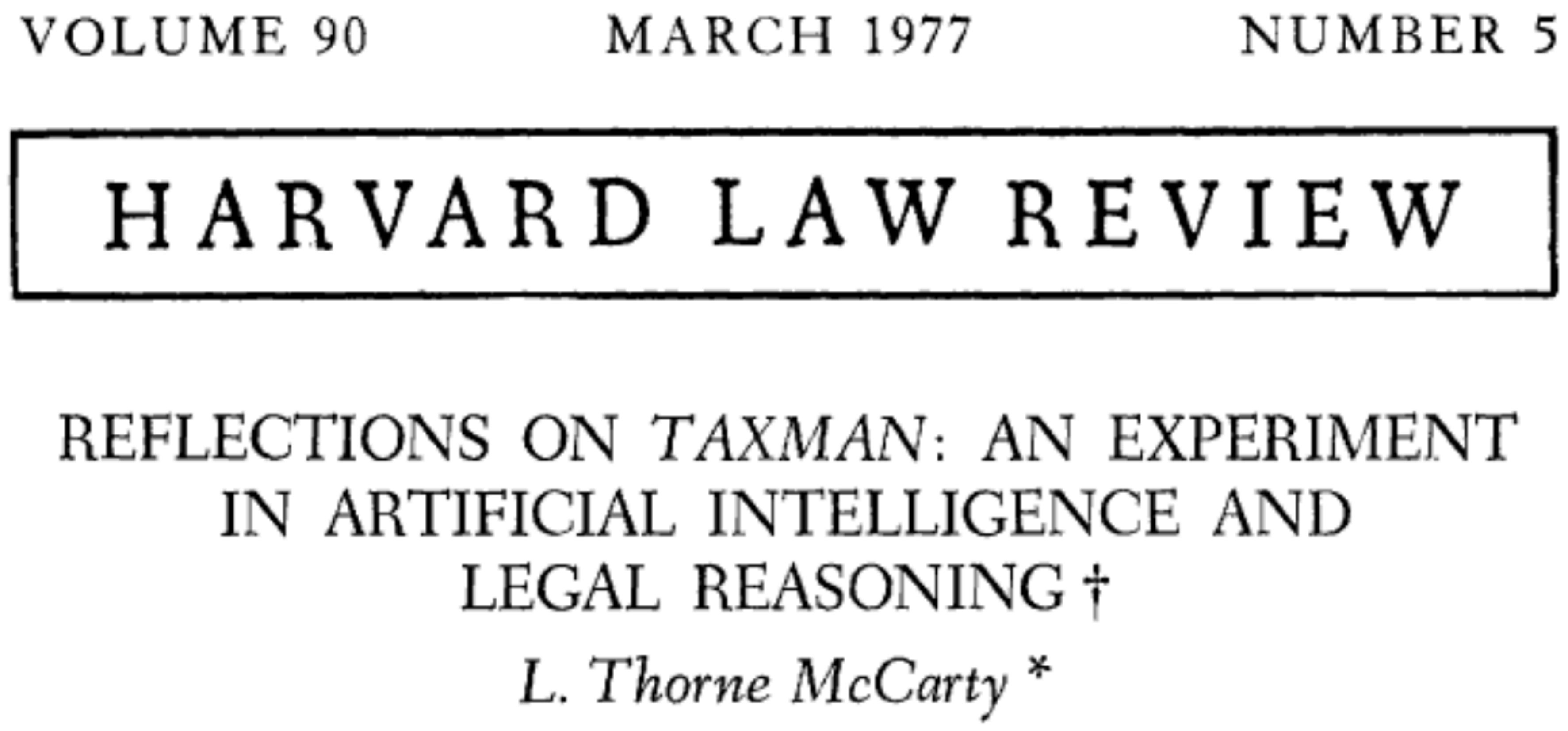
AI doesn’t even work, humans are better.

Safeguards don’t work.

**More nuance,
better regulated AI deployments,
more efficient and fair government.**

Why AI in government?

Let’s look to the IRS.



VOLUME 90 MARCH 1977 NUMBER 5

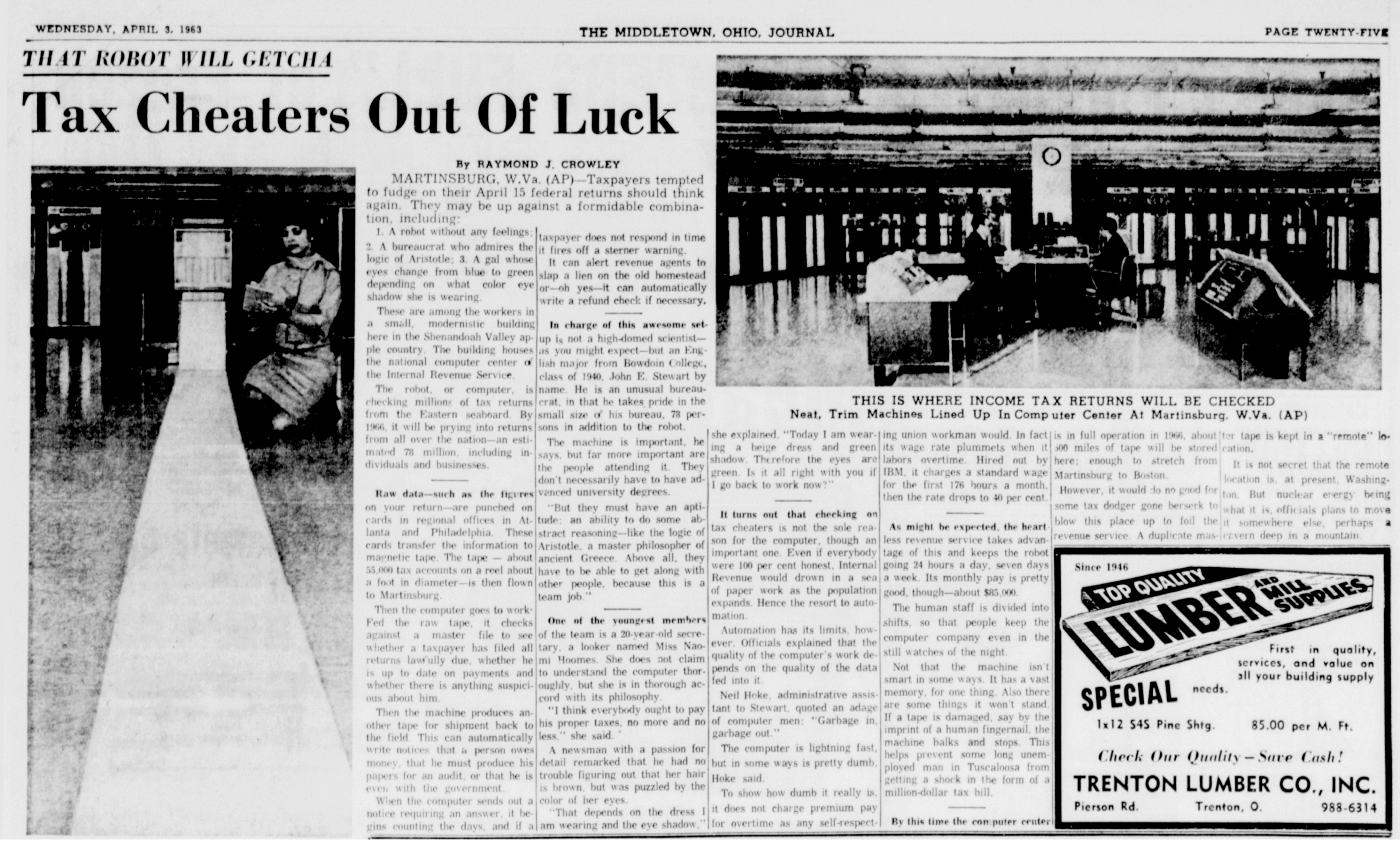
HARVARD LAW REVIEW

REFLECTIONS ON *TAXMAN*: AN EXPERIMENT IN ARTIFICIAL INTELLIGENCE AND LEGAL REASONING †

L. Thorne McCarty *

Alleged first mention of AI in a law review was related to taxes.

Let’s look to the IRS.



WEDNESDAY, APRIL 3, 1963 THE MIDDLETOWN, OHIO, JOURNAL PAGE TWENTY-FIVE

THAT ROBOT WILL GETCHA

Tax Cheaters Out Of Luck

By RAYMOND J. CROWLEY

MARTINSBURG, W.Va. (AP)—Taxpayers tempted to fudge on their April 15 federal returns should think again. They may be up against a formidable combination, including:

1. A robot without any feelings; 2. A bureaucrat who admires the logic of Aristotle; 3. A gal whose eyes change from blue to green depending on what color eye shadow she is wearing.

These are among the workers in a small, modernistic building here in the Shenandoah Valley apple country. The building houses the national computer center of the Internal Revenue Service.

The robot, or computer, is checking millions of tax returns from the Eastern seaboard. By 1966, it will be prying into returns from all over the nation—an estimated 78 million, including individuals and businesses.

Raw data—such as the figures on your return—are punched on cards in regional offices in Atlanta and Philadelphia. These cards transfer the information to magnetic tape. The tape — about 55,000 tax accounts on a reel about a foot in diameter—is then flown to Martinsburg.

Then the computer goes to work· Fed the raw tape, it checks against a master file to see whether a taxpayer has filed all returns lawfully due, whether he is up to date on payments and whether there is anything suspicious about him.

Then the machine produces another tape for shipment back to the field. This can automatically write notices that a person owes money, that he must produce his papers for an audit, or that he is even with the government.

When the computer sends out a notice requiring an answer, it begins counting the days, and if a taxpayer does not respond in time it fires off a sterner warning.

It can alert revenue agents to slap a lien on the old homestead or—oh yes—it can automatically write a refund check if necessary.

In charge of this awesome setup is not a high-domed scientist—as you might expect—but an English major from Bowdoin College, class of 1940, John E. Stewart by name. He is an unusual bureaucrat, in that he takes pride in the small size of his bureau, 78 persons in addition to the robot.

The machine is important, he says, but far more important are the people attending it. They don't necessarily have to have advanced university degrees.

"But they must have an aptitude; an ability to do some abstract reasoning—like the logic of Aristotle, a master philosopher of ancient Greece. Above all, they have to be able to get along with other people, because this is a team job."

One of the youngest members of the team is a 20-year-old secretary, a looker named Miss Naomi Hoomes. She does not claim to understand the computer thoroughly, but she is in thorough accord with its philosophy.

"I think everybody ought to pay his proper taxes, no more and no less," she said.

A newsman with a passion for detail remarked that he had no trouble figuring out that her hair is brown, but was puzzled by the color of her eyes.

"That depends on the dress I am wearing and the eye shadow," she explained. "Today I am wearing a beige dress and green shadow. Therefore the eyes are green. Is it all right with you if I go back to work now?"

It turns out that checking on tax cheaters is not the sole reason for the computer, though an important one. Even if everybody were 100 per cent honest, Internal Revenue would drown in a sea of paper work as the population expands. Hence the resort to automation.

Automation has its limits, however. Officials explained that the quality of the computer's work depends on the quality of the data fed into it.

Neil Hoke, administrative assistant to Stewart, quoted an adage of computer men: "Garbage in, garbage out."

The computer is lightning fast, but in some ways is pretty dumb, Hoke said.

To show how dumb it really is, it does not charge premium pay for overtime as any self-respecting union workman would. In fact its wage rate plummets when it labors overtime. Hired out by IBM, it charges a standard wage for the first 176 hours a month, then the rate drops to 40 per cent.

As might be expected, the heartless revenue service takes advantage of this and keeps the robot going 24 hours a day, seven days a week. Its monthly pay is pretty good, though—about $85,000.

The human staff is divided into shifts, so that people keep the computer company even in the still watches of the night.

Not that the machine isn't smart in some ways. It has a vast memory, for one thing. Also there are some things it won't stand. If a tape is damaged, say by the imprint of a human fingernail, the machine balks and stops. This helps prevent some long unemployed man in Tuscaloosa from getting a shock in the form of a million-dollar tax bill.

By this time the computer center is in full operation in 1966, about 500 miles of tape will be stored here; enough to stretch from Martinsburg to Boston.

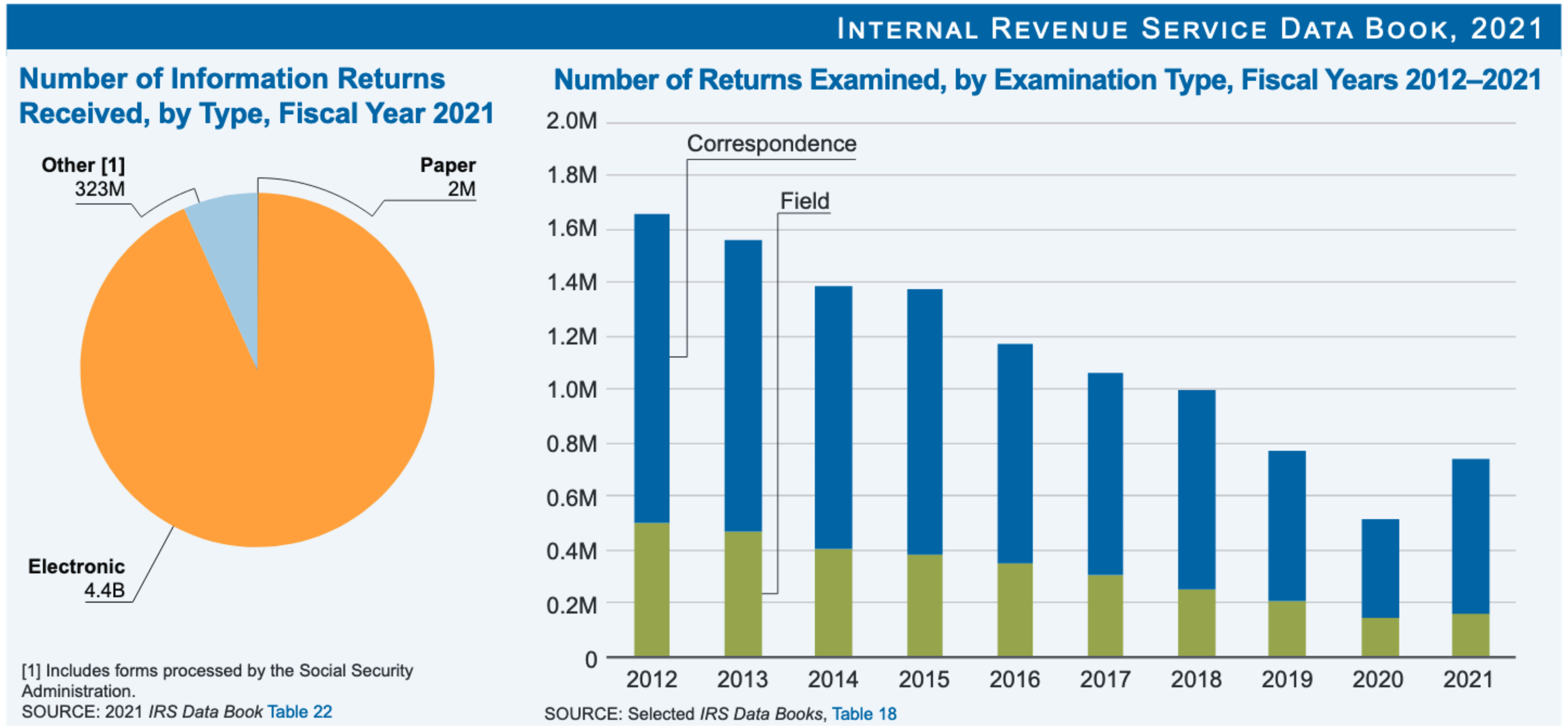
However, it would do no good for some tax dodger gone berserk to blow this place up to foil the revenue service. A duplicate master tape is kept in a "remote" location.

It is not secret that the remote location is, at present, Washington. But nuclear energy being what it is, officials plans to move it somewhere else, perhaps a cavern deep in a mountain.

THIS IS WHERE INCOME TAX RETURNS WILL BE CHECKED
Neat, Trim Machines Lined Up In Computer Center At Martinsburg, W.Va. (AP)

IRS robot in 1963.

Let's look to the IRS.



INTERNAL REVENUE SERVICE DATA BOOK, 2021

Number of Information Returns Received, by Type, Fiscal Year 2021

[1] Includes forms processed by the Social Security Administration.
SOURCE: 2021 *IRS Data Book* Table 22

Number of Returns Examined, by Examination Type, Fiscal Years 2012–2021

SOURCE: Selected *IRS Data Books*, Table 18

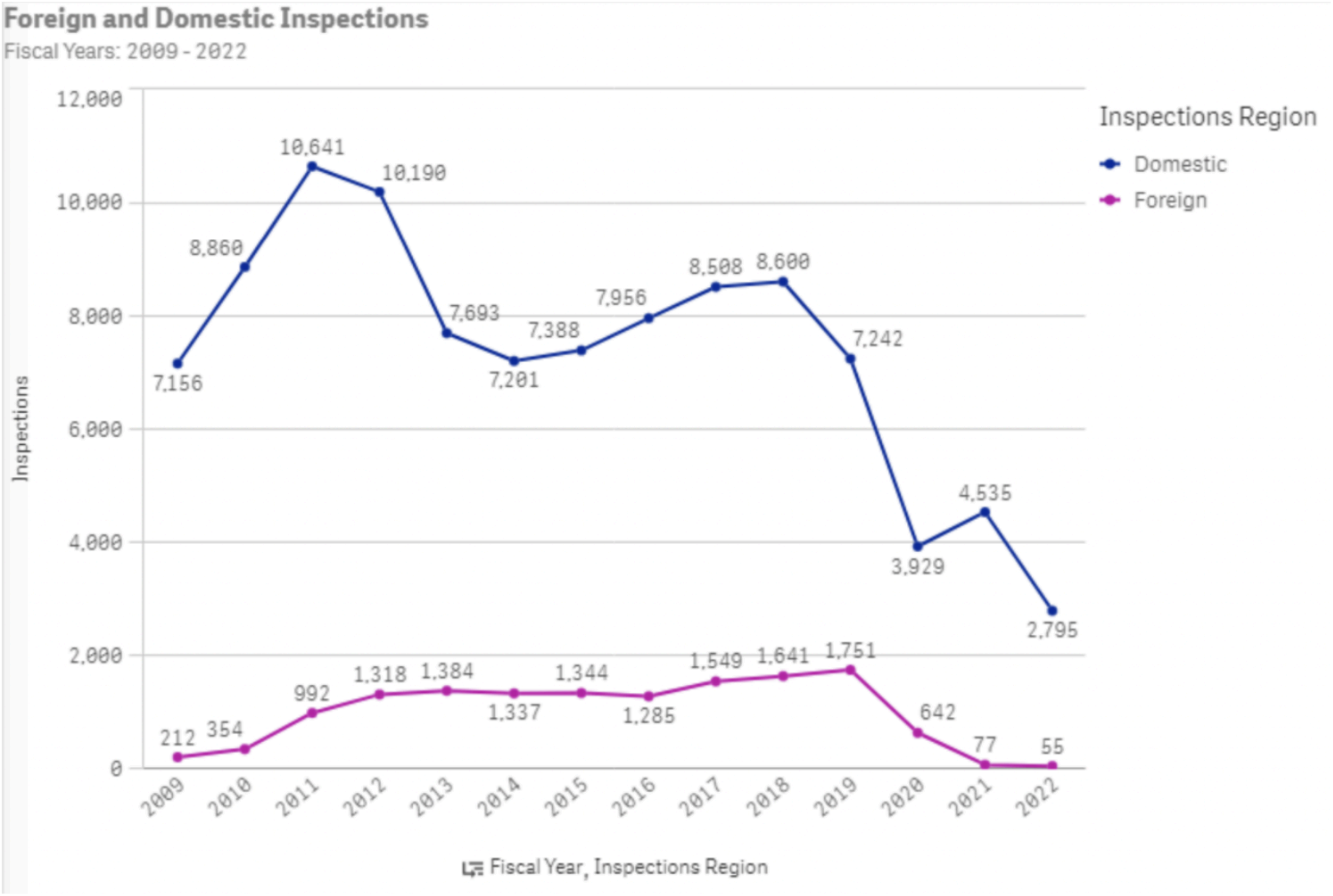
Source: IRS data book.

Source: Courtesy of Treasury Department.

Tax gap is estimated at $441 billion per year.

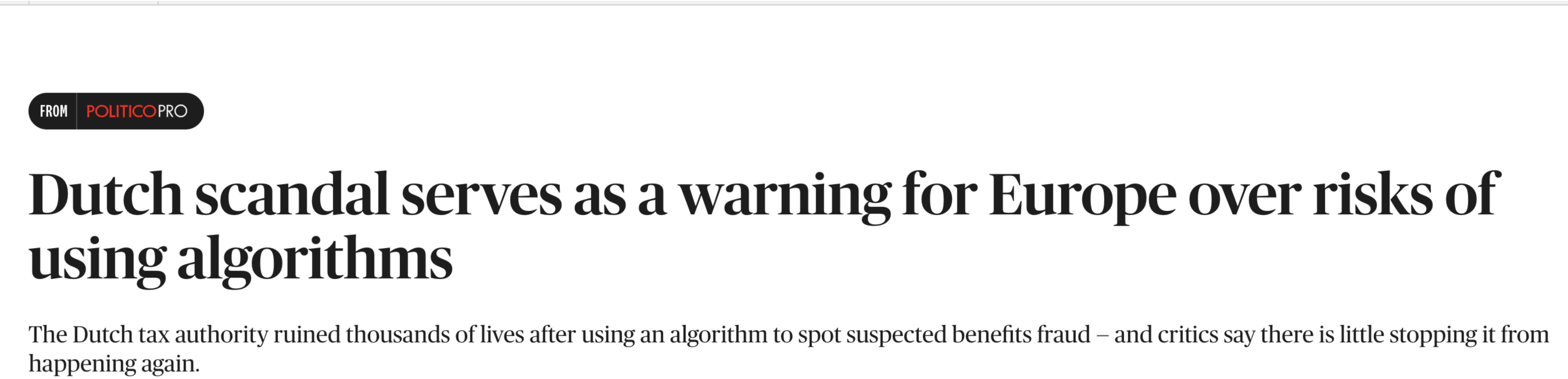
It’s nearly impossible for the IRS to do its job at this scale without smart prioritization and some forms of AI.

And this story repeats itself at other agencies that have even lower budgets for crucial government functions.



Source: Food and Drug Administration

But there are risks.



FROM POLITICOPRO

Dutch scandal serves as a warning for Europe over risks of using algorithms

The Dutch tax authority ruined thousands of lives after using an algorithm to spot suspected benefits fraud – and critics say there is little stopping it from happening again.

How do we incentivize a culture of AI Safety in gov?

How do we ensure AI Safety?

Existing laws provide some constraints and actionable lessons for AI Safety.

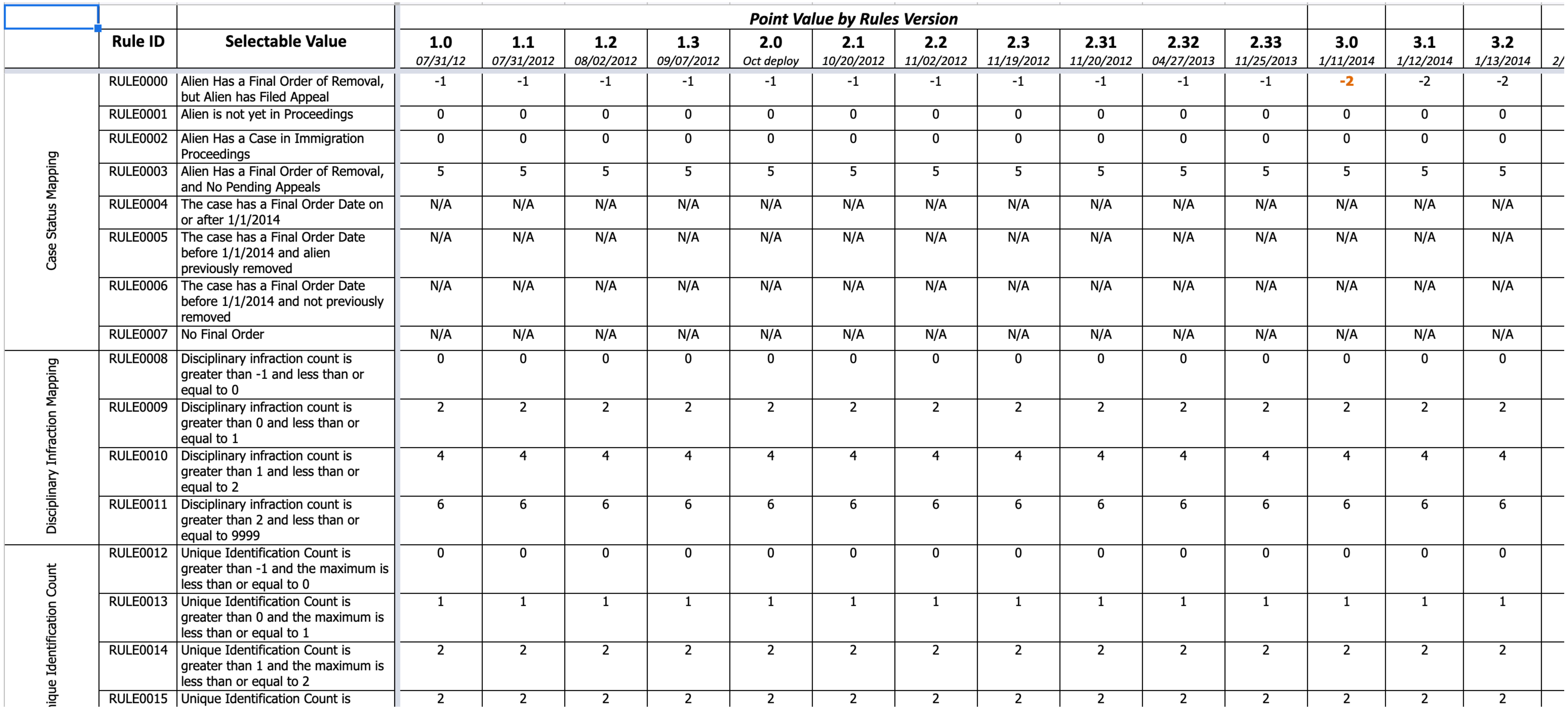
Goal: The law has something to teach AI Safety researchers and AI Safety researchers have something to teach lawmakers.

Lesson #1 from the Law:

It’s not enough for humans to just be in the loop, they have to actually be able to assert their discretion. And when they don’t, you need a fallback system that is efficient.

*From forthcoming work with Mark Krass at WeRobot 2022.

Immigration & Customs Enforcement RCA Algorithm



			Point Value by Rules Version														
	Rule ID	Selectable Value	1.0	1.1	1.2	1.3	2.0	2.1	2.2	2.3	2.31	2.32	2.33	3.0	3.1	3.2	
			07/31/12	07/31/2012	08/02/2012	09/07/2012	Oct deploy	10/20/2012	11/02/2012	11/19/2012	11/20/2012	04/27/2013	11/25/2013	1/11/2014	1/12/2014	1/13/2014	2/
Case Status Mapping	RULE0000	Alien Has a Final Order of Removal, but Alien has Filed Appeal	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-2	-2	
	RULE0001	Alien is not yet in Proceedings	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	RULE0002	Alien Has a Case in Immigration Proceedings	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	RULE0003	Alien Has a Final Order of Removal, and No Pending Appeals	5	5	5	5	5	5	5	5	5	5	5	5	5	5	
	RULE0004	The case has a Final Order Date on or after 1/1/2014	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	RULE0005	The case has a Final Order Date before 1/1/2014 and alien previously removed	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	RULE0006	The case has a Final Order Date before 1/1/2014 and not previously removed	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	RULE0007	No Final Order	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
Disciplinary Infraction Mapping	RULE0008	Disciplinary infraction count is greater than -1 and less than or equal to 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	RULE0009	Disciplinary infraction count is greater than 0 and less than or equal to 1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
	RULE0010	Disciplinary infraction count is greater than 1 and less than or equal to 2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	
	RULE0011	Disciplinary infraction count is greater than 2 and less than or equal to 9999	6	6	6	6	6	6	6	6	6	6	6	6	6	6	
ique Identification Count	RULE0012	Unique Identification Count is greater than -1 and the maximum is less than or equal to 0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	RULE0013	Unique Identification Count is greater than 0 and the maximum is less than or equal to 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	RULE0014	Unique Identification Count is greater than 1 and the maximum is less than or equal to 2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	
	RULE0015	Unique Identification Count is	2	2	2	2	2	2	2	2	2	2	2	2	2	2	

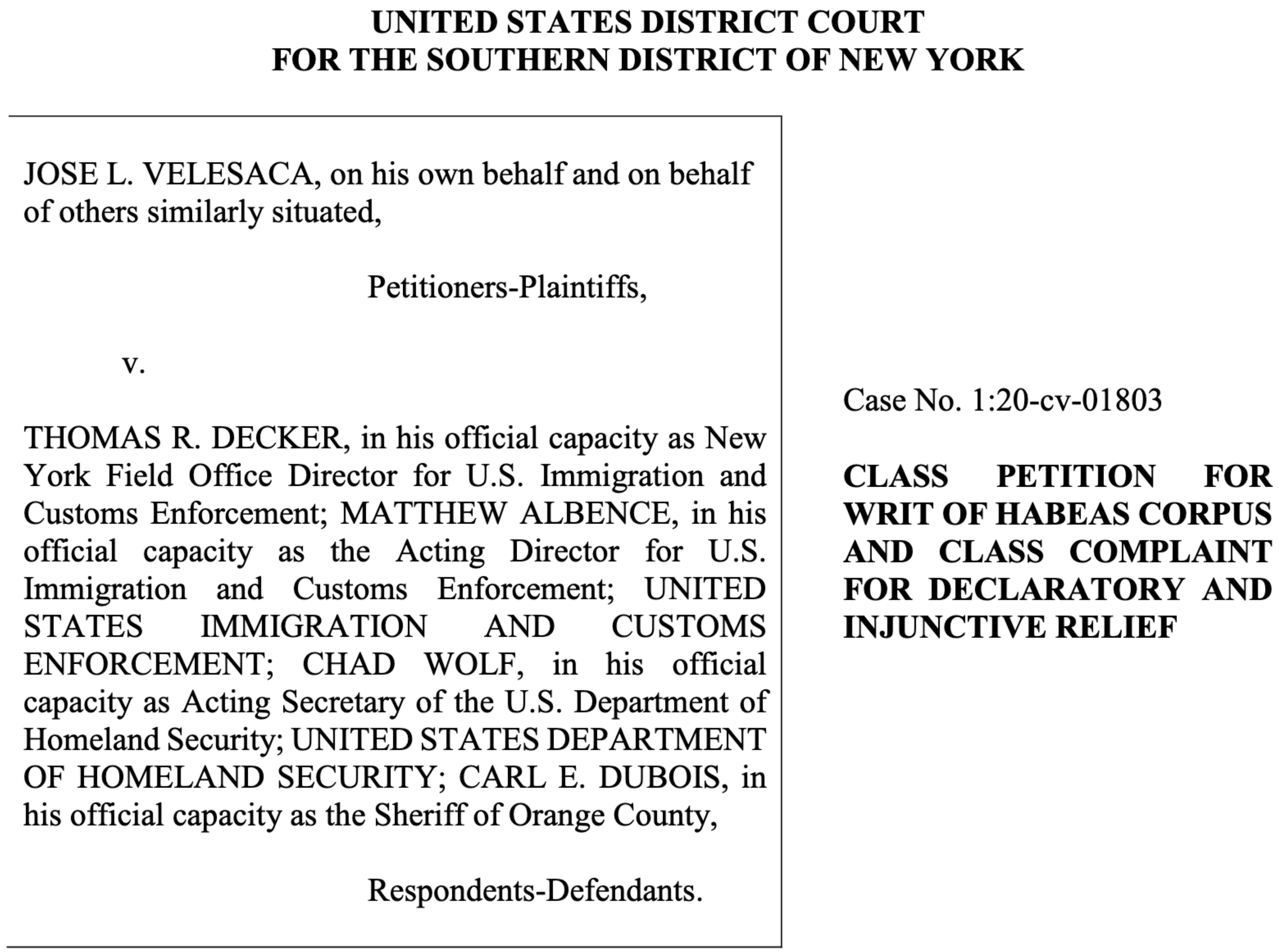
Immigration & Customs Enforcement RCA Algorithm

Similar story to Dutch Tax Service.

Officers began to rely on algorithm for recommendations, and stopped having discretion.

Eventually, algorithm was silently changed to never allow release for anyone.

Immigration & Customs Enforcement RCA Algorithm



UNITED STATES DISTRICT COURT
FOR THE SOUTHERN DISTRICT OF NEW YORK

JOSE L. VELESACA, on his own behalf and on behalf of others similarly situated,

Petitioners-Plaintiffs,

v.

THOMAS R. DECKER, in his official capacity as New York Field Office Director for U.S. Immigration and Customs Enforcement; MATTHEW ALBENCE, in his official capacity as the Acting Director for U.S. Immigration and Customs Enforcement; UNITED STATES IMMIGRATION AND CUSTOMS ENFORCEMENT; CHAD WOLF, in his official capacity as Acting Secretary of the U.S. Department of Homeland Security; UNITED STATES DEPARTMENT OF HOMELAND SECURITY; CARL E. DUBOIS, in his official capacity as the Sheriff of Orange County,

Respondents-Defendants.

Case No. 1:20-cv-01803

CLASS PETITION FOR WRIT OF HABEAS CORPUS AND CLASS COMPLAINT FOR DECLARATORY AND INJUNCTIVE RELIEF

Judge allowed injunction *because officers are required to exercise discretion* (among other reasons) and as a result should have gone through rule making process, which requires notice-and-comment period.

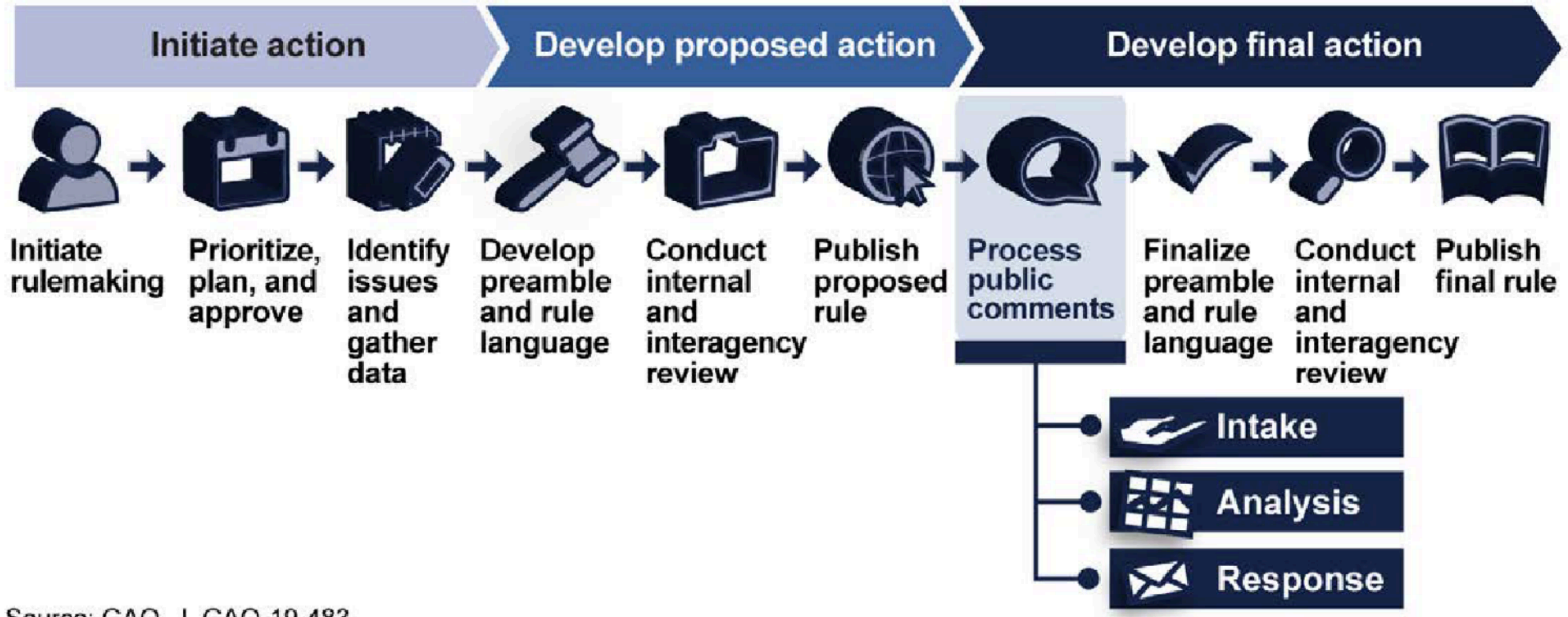
Immigration & Customs Enforcement RCA Algorithm

Demonstrates a procedural mechanism for *requiring attentive humans in the loop* *by law.*

Can teach us how to build AI Safety systems that align with administrative law.

But, there are problems. If courts require rulemaking, it can be quite long and arduous. It is not suitable for safely iterating and updating AI algorithms.

The Rulemaking Process under the Administrative Procedure Act



Initiate action → Develop proposed action → Develop final action

Initiate rulemaking → Prioritize, plan, and approve → Identify issues and gather data → Develop preamble and rule language → Conduct internal and interagency review → Publish proposed rule → Process public comments → Finalize preamble and rule language → Conduct internal and interagency review → Publish final rule

Process public comments:
- Intake
- Analysis
- Response

Source: GAO. | GAO-19-483

Lesson #1 from the Law:

It’s not enough for humans to just be in the loop, they have to actually be able to assert their discretion. And when they don’t, you need a fallback system that is efficient.

Courts and Administrative Agencies Balance Transparency against Privacy

Lesson #2 from the Law:

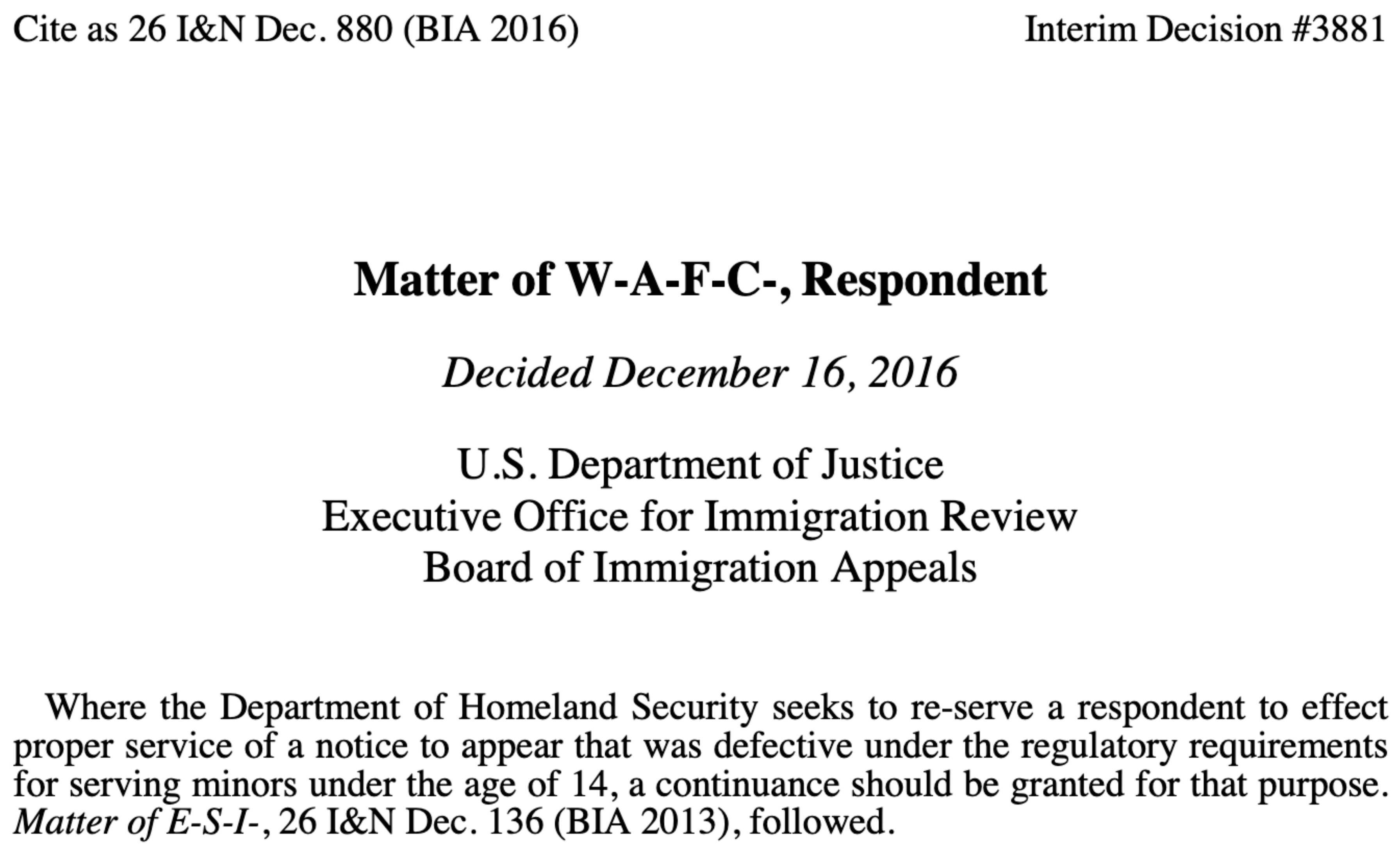
Transparency and openness is key to fight corruption and ensure safety.
But you have to find ways to balance that against privacy interests in a highly contextual way.

*From Peter Henderson, Mark Simon Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. "Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset." (2022).

Courts and Administrative Agencies Balance Transparency against Privacy

Courts and agencies want to (and actually *have to*) release their decisions and detailed reasoning for them. But this necessarily means including personal details about the situation under discussion.

Courts and Administrative Agencies Balance Transparency against Privacy



Cite as 26 I&N Dec. 880 (BIA 2016) Interim Decision #3881

Matter of W-A-F-C-, Respondent

Decided December 16, 2016

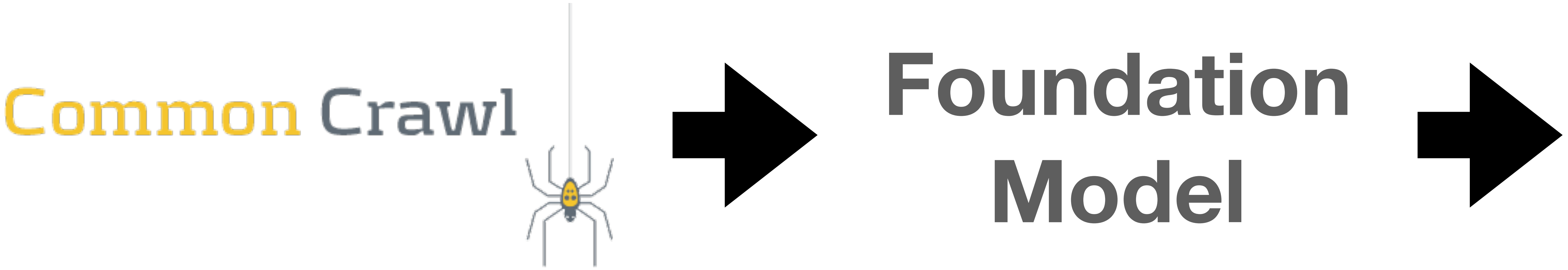
U.S. Department of Justice
Executive Office for Immigration Review
Board of Immigration Appeals

Where the Department of Homeland Security seeks to re-serve a respondent to effect proper service of a notice to appear that was defective under the regulatory requirements for serving minors under the age of 14, a continuance should be granted for that purpose. *Matter of E-S-I-*, 26 I&N Dec. 136 (BIA 2013), followed.

One way is to use pseuodonyms and to redact enough information so the person cannot be identified.

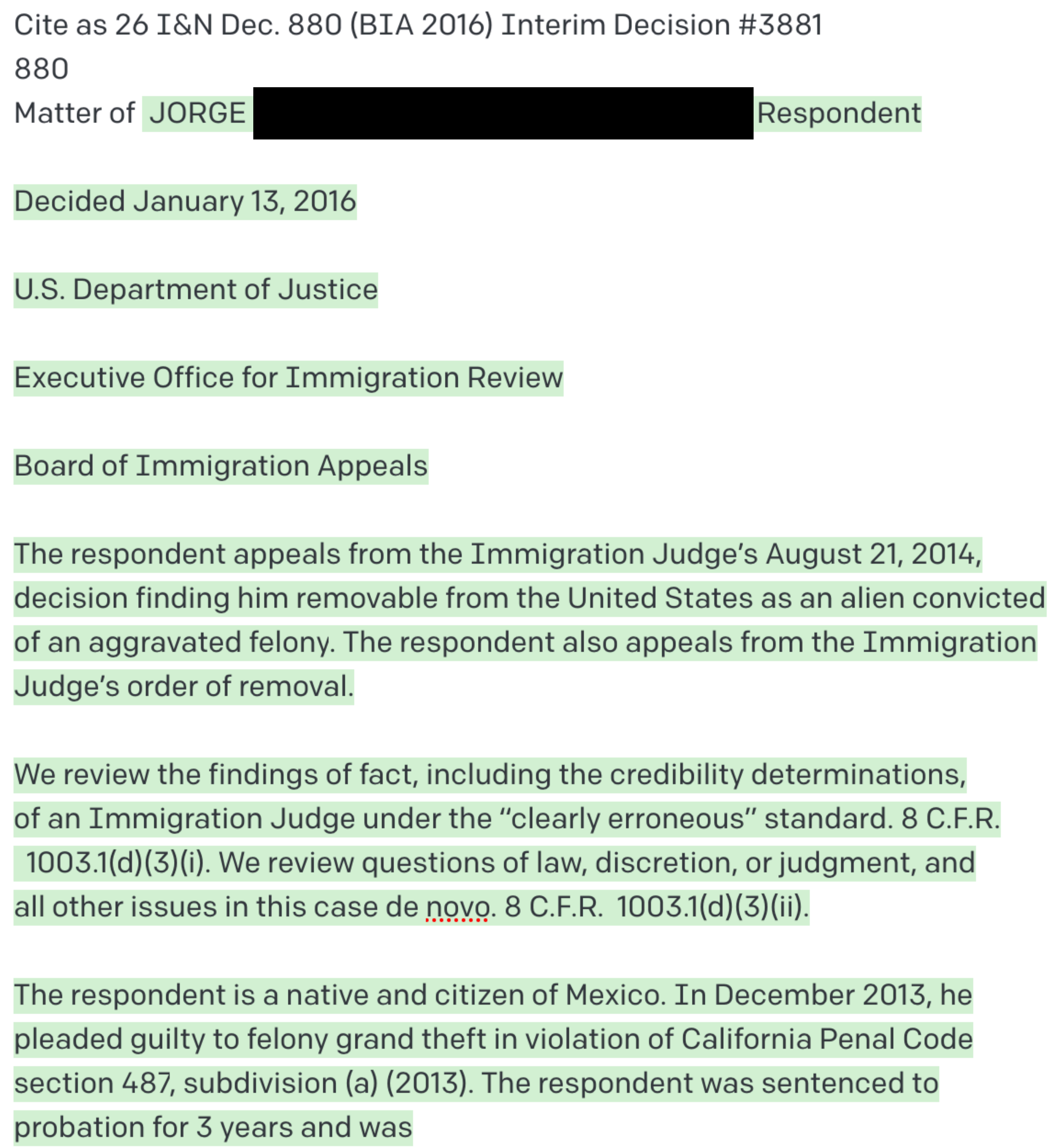
Courts and Administrative Agencies Balance Transparency against Privacy

Compare that against what we do for large language models.



Common Crawl

Foundation Model



Cite as 26 I&N Dec. 880 (BIA 2016) Interim Decision #3881
880
Matter of JORGE [redacted] Respondent

Decided January 13, 2016

U.S. Department of Justice

Executive Office for Immigration Review

Board of Immigration Appeals

The respondent appeals from the Immigration Judge's August 21, 2014, decision finding him removable from the United States as an alien convicted of an aggravated felony. The respondent also appeals from the Immigration Judge's order of removal.

We review the findings of fact, including the credibility determinations, of an Immigration Judge under the "clearly erroneous" standard. 8 C.F.R. 1003.1(d)(3)(i). We review questions of law, discretion, or judgment, and all other issues in this case de novo. 8 C.F.R. 1003.1(d)(3)(ii).

The respondent is a native and citizen of Mexico. In December 2013, he pleaded guilty to felony grand theft in violation of California Penal Code section 487, subdivision (a) (2013). The respondent was sentenced to probation for 3 years and was

Courts and Administrative Agencies Balance Transparency against Privacy

Unclear if generated content is safe to release.

People's names might be associated with information that might cause safety harms.

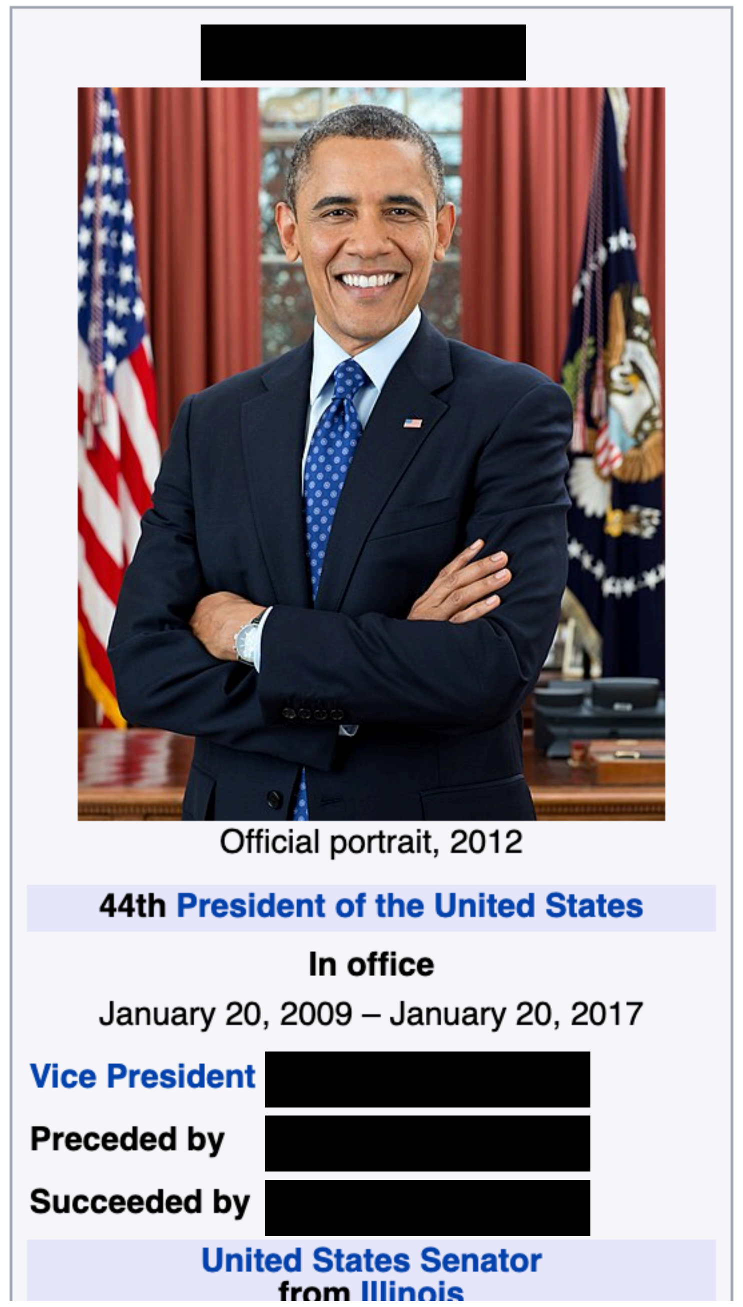
The information would have to be out on the web already, but sometimes it is harder to find (de-indexed from Google, etc.).

Models don't respect this.

Table 1: Filters Applied in Major Pre-Training Papers

	PSI	Deduplication	Toxic Content	Quality
CCNet [122]	No	MinHash (pages)	No	No
C4 [96]	No	Unknown (3-sentence spans)	Word list	Minimum word counts, presence of curly brackets, ‘lorem ipsum’, etc.
GPT-3 [21]	No	MinHash (pages)	No	Train classifier to distinguish CC from curated high-quality examples
Gopher [95]	No	MinHash (pages)	Google Safe-Search	Min./max. word counts, word-to-symbol ratio, share ellipses, excessive repetition; require stop words
The Pile [44]	No	MinHash (pages)	Ad-hoc source deletion	Train classifier to distinguish CC from curated high-quality examples

Courts and Administrative Agencies Balance Transparency against Privacy



[redacted]

Official portrait, 2012

44th President of the United States

In office

January 20, 2009 – January 20, 2017

Vice President	[redacted]
Preceded by	[redacted]
Succeeded by	[redacted]

United States Senator
from Illinois

Who is the 44th President of the United States?

Redacted Model: ???
Unredacted Model: Barack Obama

How do we redact names in situations that might be unsafe, but keep names in situations where it's necessary.

For example, case names are laws in common law systems, cannot redact. Or you might want to retain information about public figures or characters in a movie.

Courts and Administrative Agencies Balance Transparency against Privacy

The law can teach us (imperfectly)! Executive Office of Immigration Review and other agencies make these decisions daily.

Table 2: Availability of Identifying Information Across Administrative Settings

Jurisdiction	Civil Cases	Criminal Cases	Juvenile Data
U.S. Federal Courts	All case details public unless sealed, except DOBs, ID/account #s.	Def. names public; DOBs, ID/account #s, addresses redacted.	Criminal records confidential. Names redacted from civil cases.
U.S. Admin. Agencies	Most PII omitted from public records.	-	No statute; more protection in practice.
German Courts	Judgments omit all identifying information.	Confidential 3-5 years after sentence completed.	No public access to criminal records.
Chinese Courts	Names/case details public except in certain classes of cases.	Names/case details are public as of 2016.	Juvenile criminal records are categorically exempt from disclosure.
Canadian Courts	Presumption of openness, except specific details and rare sealed cases.	Public; may be sealed after a period of good behavior.	Youth criminal records are always confidential.

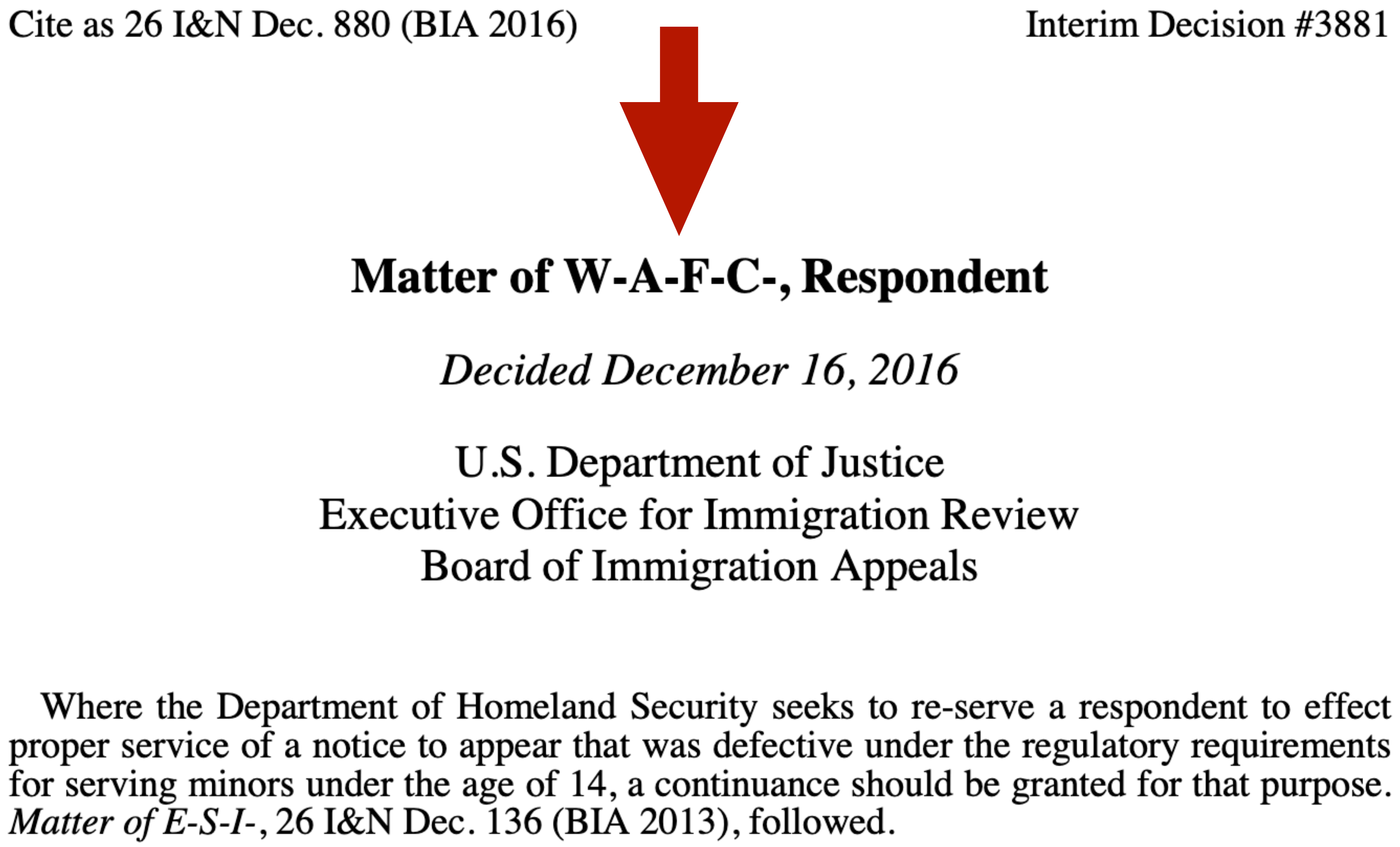
Table 4: Description of the Pile of Law by Data Source

Data Source	Data Size	Word Count	Document Count
Court Listener Opinions	59.29GB/19.76GB	7.65B/2.55B	3.39M/1.12M
Court Listener Docket Entries and Court Filings	52.13GB/17.38GB	5.36B/1.79B	1.49M/496K
U.S. Supreme Docket Entries and Court Filings	1.51GB/0.50GB	151.05M/51.73M	48K/16K
U.S. Board of Veterans' Appeals Decisions	13.21GB/4.40GB	1.74B/580.98M	630K/210K
U.S. Federal Trade Commission Advisory Opinions	1.55MB/0.52MB	157K/53K	112/33
U.S. National Labor Relations Board Decisions	994.83MB/331.61MB	120.33M/39.20M	24K/8K
U.S. Department of Justice Executive Office for Immigration Review *Immigration & Nationality Decisions*	22.89MB/7.63MB	3.05M/1.01M	1671/558
U.S. Department of Labor Employees' Compensation Appeals Board	353.25MB/117.75MB	48.20M/16.01M	21K/7K
European Court of Human Rights Opinions [91]	111.53MB/37.18MB	16.71M/3.47M	7K/1K
Canadian Court Opinions (ON, BC)	182.09MB/60.70MB	23.45M/7.66M	8K/3K
U.S. Office of Legal Counsel Memos	36.98MB/12.33MB	4.36M/1.31M	1038/346
U.S. Office of Inspector General Reports	1.90GB/0.63GB	167.71M/54.18M	29K/10K
U.S. Code of Federal Regulations	670.87MB/223.62MB	79.06M/25.41M	182/61
U.S. Supreme Court Oral Argument Transcripts	1.51GB/0.50GB	151.05M/51.73M	47K/16K
U.S. State Codes	6.77GB/2.26GB	829.62M/441.38M	157/60
U.S. Code	268.40MB/89.47MB	30.54M/18.20M	43/15
U.S. Federal Rules of Evidence	670KB/223KB	77K/36K	51/17
U.S. Federal Rules of Civil Procedure	1.59MB/0.53MB	237K/40K	69/23
U.S. Bills	1.27GB/0.42GB	156.06M/49.4M	84K/28K
U.S. Federal Register	159.29MB/53.10MB	6.61M/53.27M	4060/1354
U.S. Founders Letters	419.33MB/139.78MB	53.27M/17.69M	138K/46K
World Constitutions [41]	24.43MB/8.14MB	3.43M/1.06M	139/48
EUR-Lex [28]	1.31GB/0.44GB	191.65M/65.31M	106K/36K
Credit Card Agreements	70.19MB/23.40MB	10.73M/3.09M	2023/615
Terms of Service [75, 99]	1.57MB/0.52MB	213K/62K	37/13
Edgar Contracts [17]	10.76GB/3.59GB	1.44B/473.50M	741K/247K
Atticus Contracts [55]	31.2GB/10.4GB	3.96B/1.31B	488K/163K
U.S. Congressional Hearings	6.17GB/2.06GB	761.12M/250.04M	24K/8K
U.S. Tax Court PLR Corpus [14]	639.03MB/213.01MB	84.25M/27.62M	41K/14K
European Parliament Proceedings Parallel Corpus [63]	302.71MB/100.90MB	41.55M/13.70M	7K/2K
U.N. General Debate Corpus [8]	134.90MB/44.97MB	17.68M/5.81M	6K/2K
Reddit r/legaladvice & r/legaladviceofftopic	299.04MB/99.68MB	40.42M/13.56M	110K/37K
Bar Exam Outlines	1.18MB/0.39MB	123K/43K	44/15
Open Source Casebooks	87.09MB/29.03MB	9.20M/3.91M	52/14
Total	$\sim$ 256GB	$\sim$ 30B	$\sim$ 10M

Table 5: Filtering Norms by Data Source in the Pile of Law

Data Source	Examples of Filtering Norms
Court Listener Opinions	FRCP 49.1 (requiring partial redactions for social-security number and taxpayer-identification number, date of birth, minor's names, financial account numbers; governing sealing and redaction standards for other information that parties may wish to keep private); State Rules for filing pseudonymously[9]. Judicial code of ethics govern conduct of judges; American Bar Association Model Rules of Professional Conduct govern attorney conduct.
Court Listener Docket Entries and Court Filings	*Id.*
U.S. Supreme Docket Entries and Court Filings	*Id.*
U.S. Board of Veterans' Appeals Decisions	38 CFR 20.1301(c) ("Appeals on or after January 1, 1992, are electronically available for public inspection and copying on the Board's website. All personal identifiers are redacted from the decisions prior to publication.")

[9] https://withoutmyconsent.org/50state/filing-pseudonymously/federal/



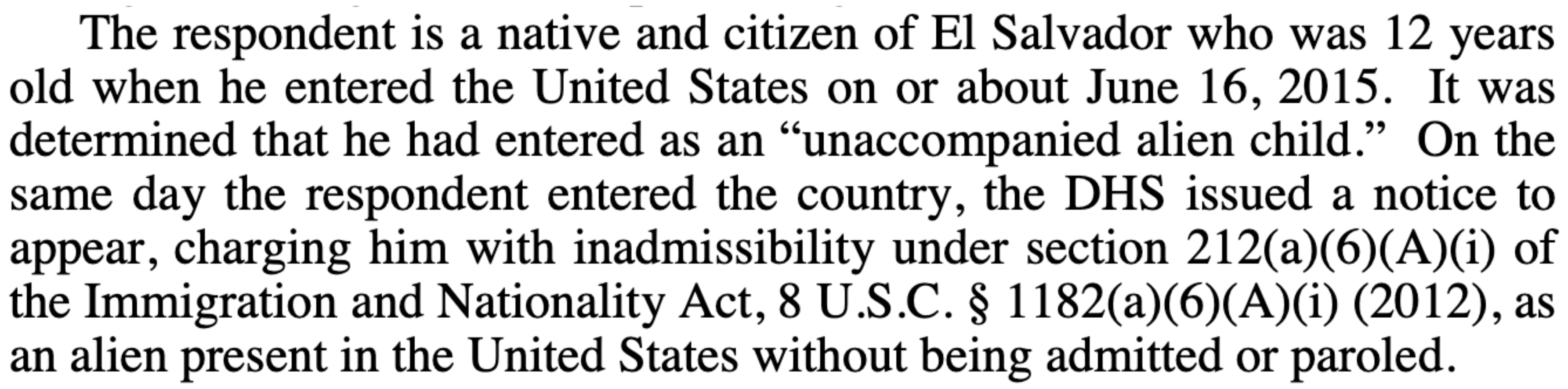
Cite as 26 I&N Dec. 880 (BIA 2016) Interim Decision #3881

Matter of W-A-F-C-, Respondent

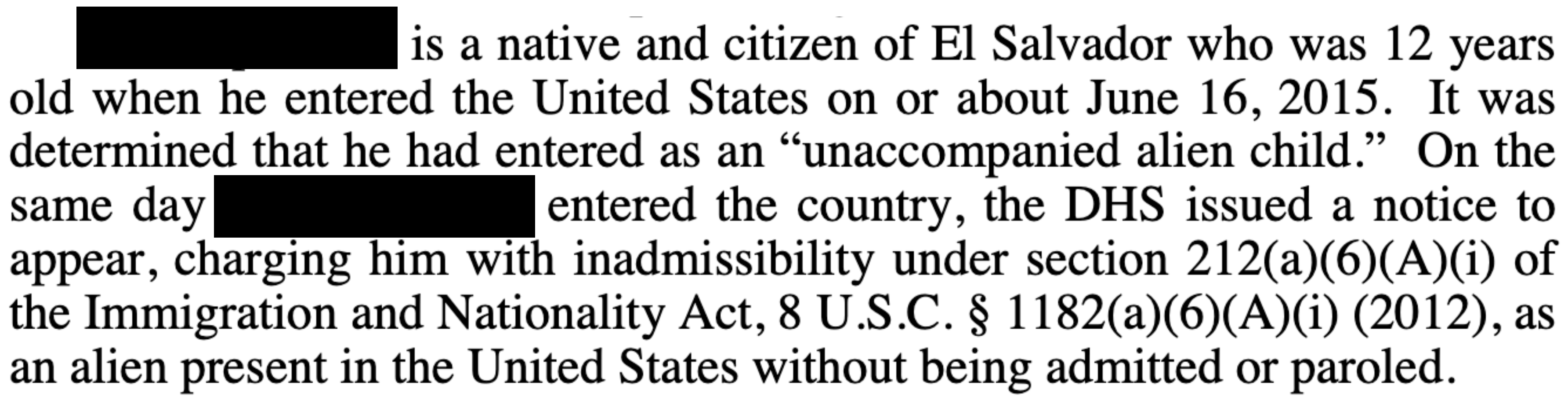
Decided December 16, 2016

U.S. Department of Justice
Executive Office for Immigration Review
Board of Immigration Appeals

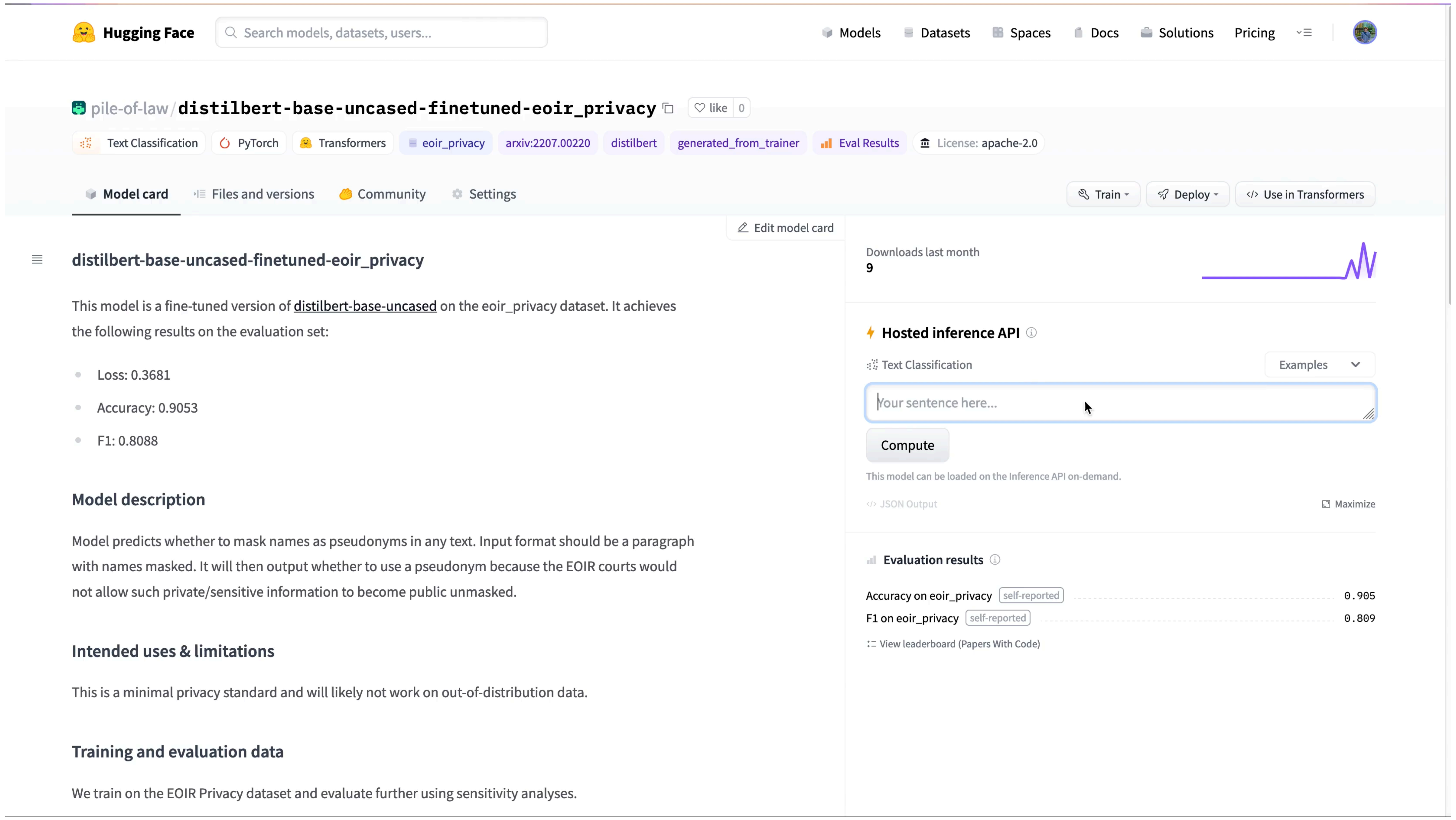
Where the Department of Homeland Security seeks to re-serve a respondent to effect proper service of a notice to appear that was defective under the regulatory requirements for serving minors under the age of 14, a continuance should be granted for that purpose. *Matter of E-S-I-*, 26 I&N Dec. 136 (BIA 2013), followed.



The respondent is a native and citizen of El Salvador who was 12 years old when he entered the United States on or about June 16, 2015. It was determined that he had entered as an "unaccompanied alien child." On the same day the respondent entered the country, the DHS issued a notice to appear, charging him with inadmissibility under section 212(a)(6)(A)(i) of the Immigration and Nationality Act, 8 U.S.C. § 1182(a)(6)(A)(i) (2012), as an alien present in the United States without being admitted or paroled.



[redacted] is a native and citizen of El Salvador who was 12 years old when he entered the United States on or about June 16, 2015. It was determined that he had entered as an "unaccompanied alien child." On the same day [redacted] entered the country, the DHS issued a notice to appear, charging him with inadmissibility under section 212(a)(6)(A)(i) of the Immigration and Nationality Act, 8 U.S.C. § 1182(a)(6)(A)(i) (2012), as an alien present in the United States without being admitted or paroled.



distilbert-base-uncased-finetuned-eoir_privacy

This model is a fine-tuned version of distilbert-base-uncased on the eoir_privacy dataset. It achieves the following results on the evaluation set:

- Loss: 0.3681
- Accuracy: 0.9053
- F1: 0.8088

Model description

Model predicts whether to mask names as pseudonyms in any text. Input format should be a paragraph with names masked. It will then output whether to use a pseudonym because the EOIR courts would not allow such private/sensitive information to become public unmasked.

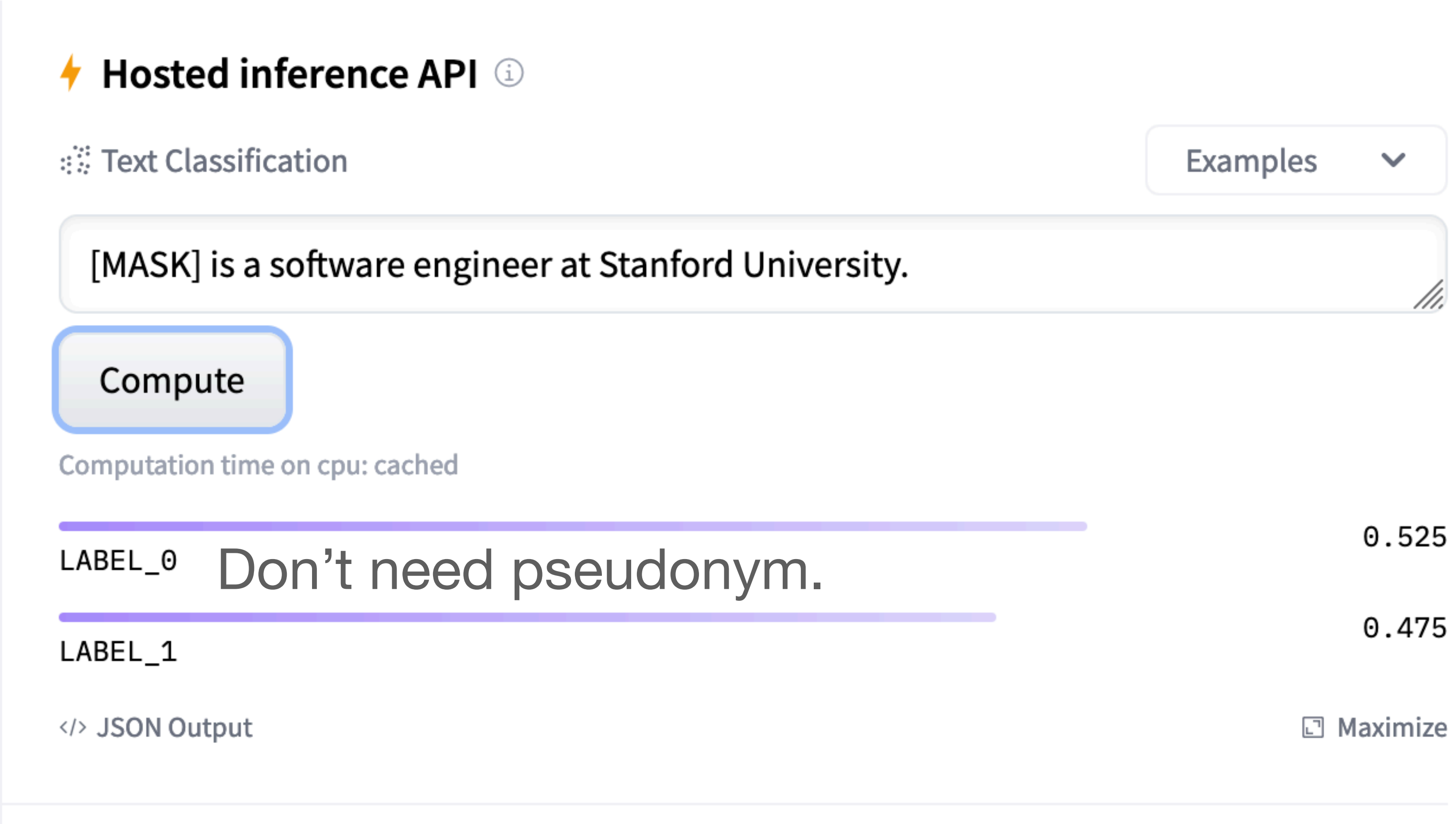
Intended uses & limitations

This is a minimal privacy standard and will likely not work on out-of-distribution data.

Training and evaluation data

We train on the EOIR Privacy dataset and evaluate further using sensitivity analyses.

Accuracy on eoir_privacy	self-reported	0.905
F1 on eoir_privacy	self-reported	0.809

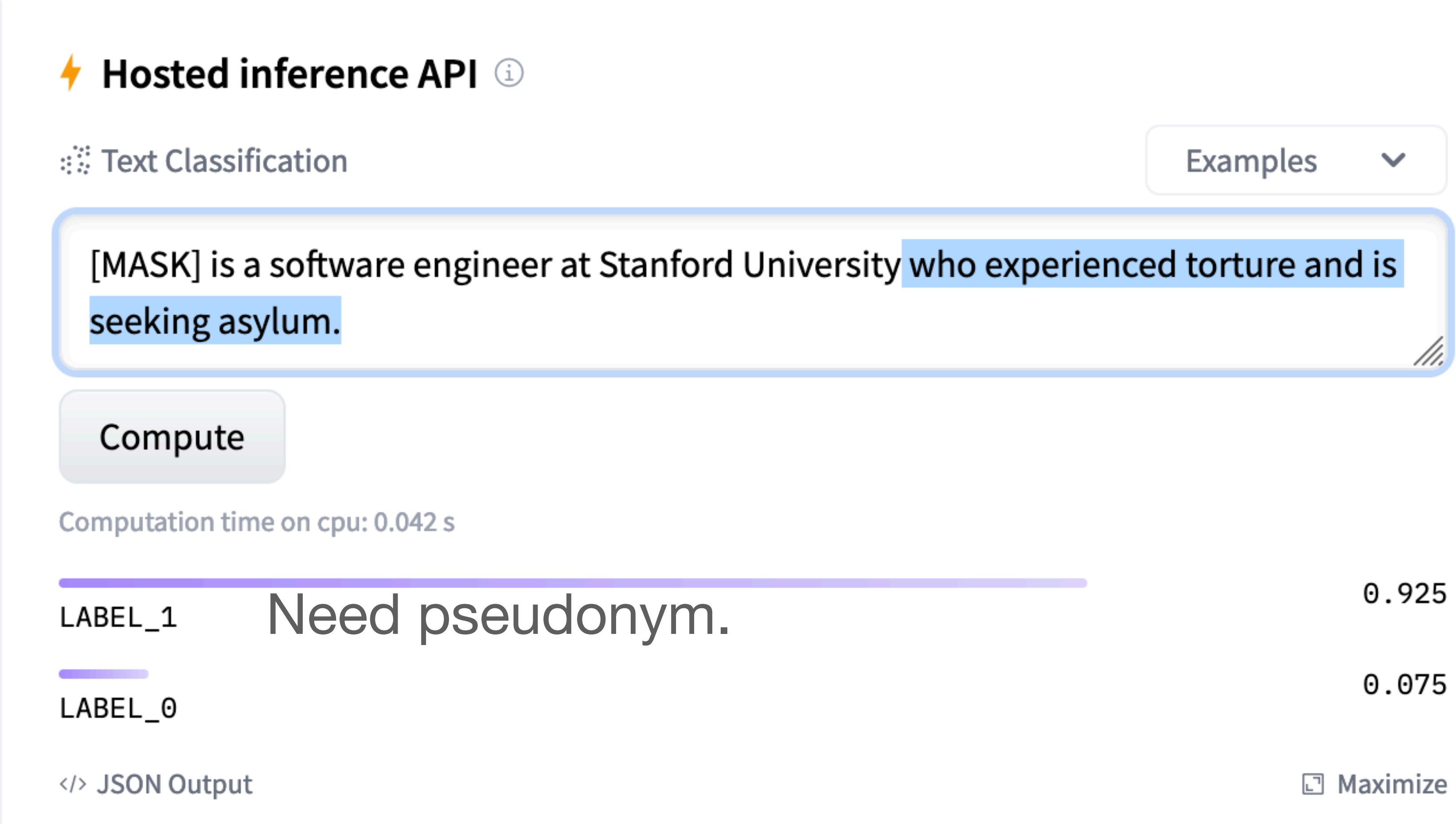


[MASK] is a software engineer at Stanford University.

Computation time on cpu: cached

```
LABEL_0   0.525
LABEL_1   0.475
```

Don’t need pseudonym.



[MASK] is a software engineer at Stanford University who experienced torture and is seeking asylum.

Computation time on cpu: 0.042 s

```
LABEL_1   0.925
LABEL_0   0.075
```

Need pseudonym.

Courts and Administrative Agencies Balance Transparency against Privacy

Lesson #2 from the Law:

Transparency and openness is key to fight corruption and ensure safety.
But you have to find ways to balance that against privacy interests in a highly contextual way.

*From Peter Henderson, Mark Simon Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. "Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset." (2022).

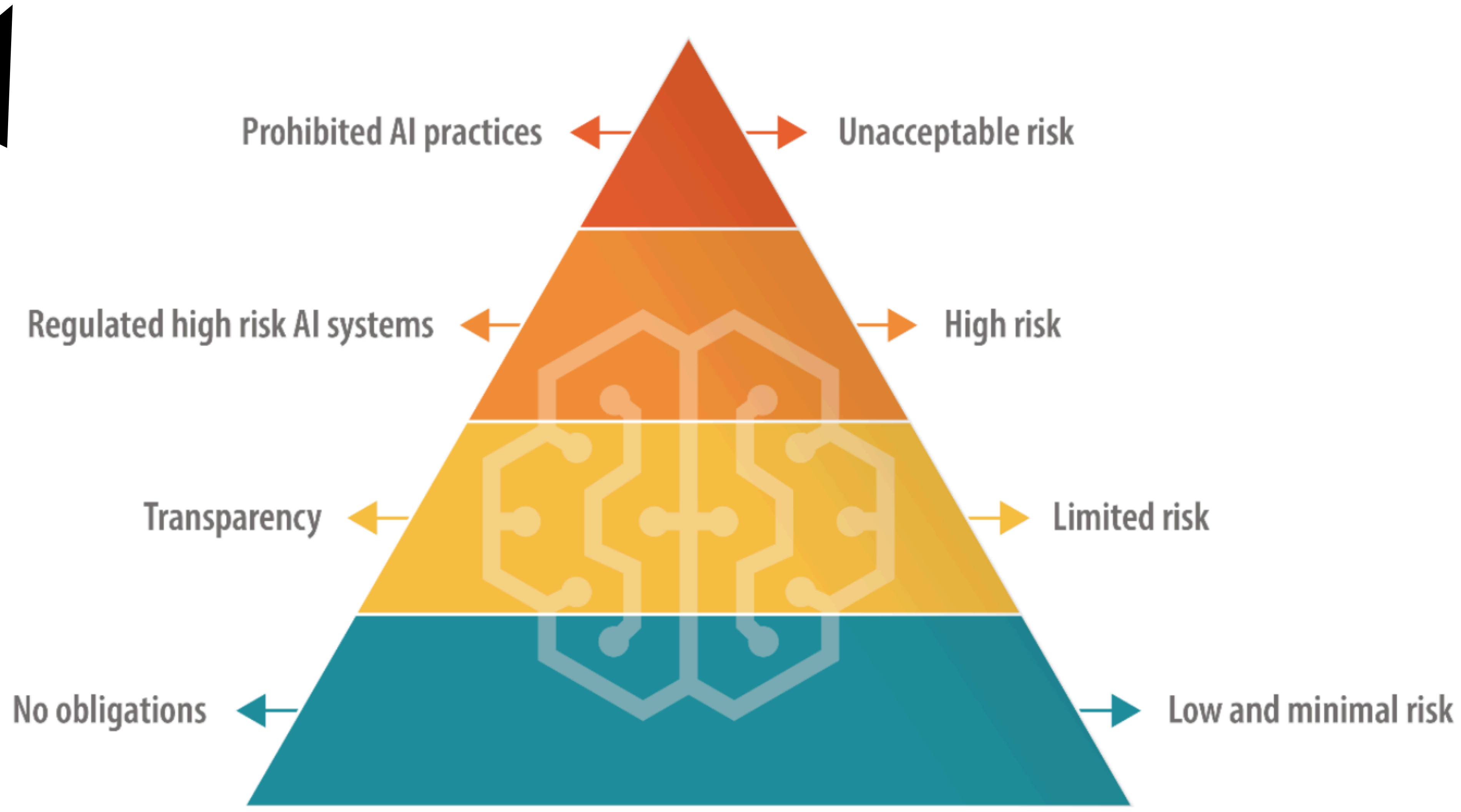
I could go on with more lessons.

But the point is that the law and AI safety are deeply intertwined, especially when you look at the constraints placed on the U.S. government.

And this might also give you some thoughts on how we might want to think about regulation for the private sector.

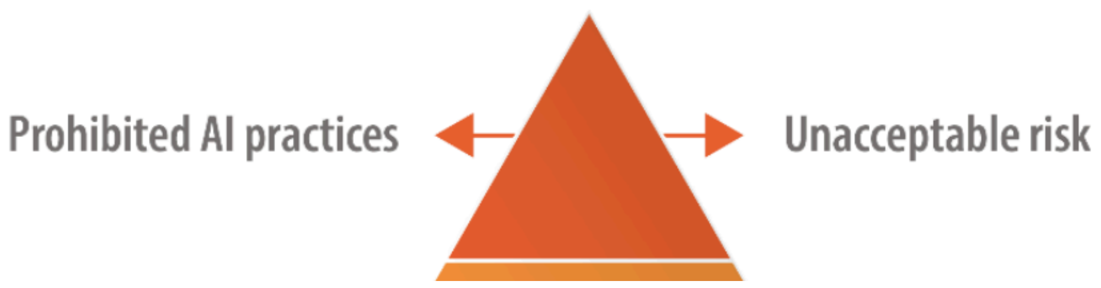
In fact, the EU AI Act does something like this.

More like sensitive government (especially autocratic government uses).



Data source: European Commission.

Less like government uses. (e.g., Generative art)



Prohibited AI practices

Unacceptable risk

Bans:

- **Any system that deploys harmful manipulative “subliminal techniques”**
- **AI systems that exploit specific vulnerable groups**
- **AI systems used by authorities for social scoring**
- **“Real-time” remote biometric ID in publicly accessible areas for law enforcement purposes.**

Transparency, Monitoring, and *ex-ante* Assessments:
Remind you of rule-making?

Regulated high risk AI systems

High risk

- Biometric identification and categorisation of natural persons;
- Management and operation of critical infrastructure;
- Education and vocational training;
- Employment, worker management and access to self-employment;
- Access to and enjoyment of essential private services and public services and benefits;
- Law enforcement;
- Migration, asylum and border control management;
- Administration of justice and democratic processes.

We need to get into both a technical and a regulatory law mindset to make AI Safety well-formed.

Feel free to reach out!