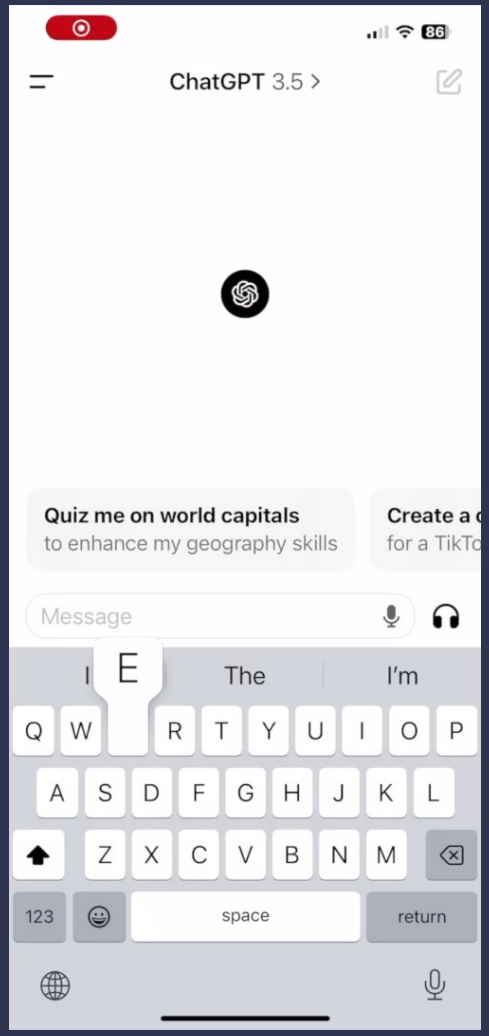
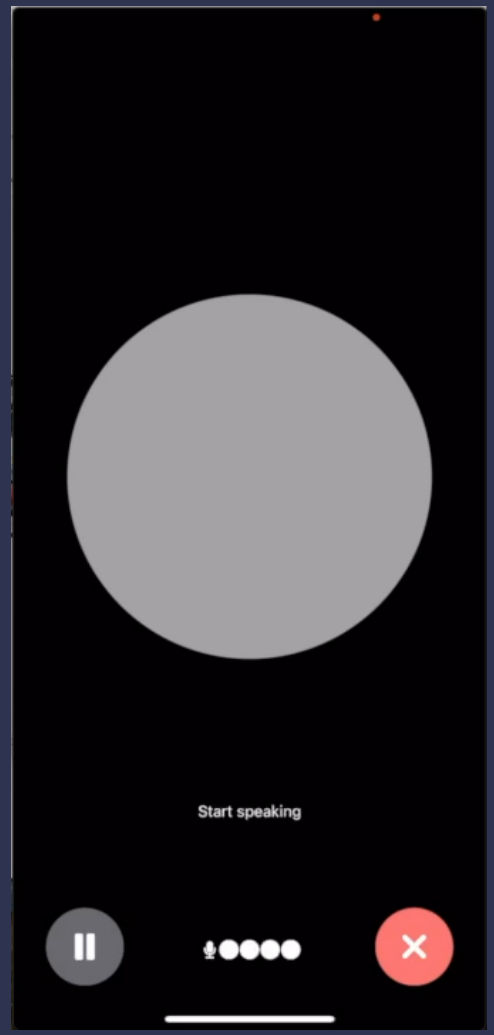
Ethical and Safety Considerations in AI

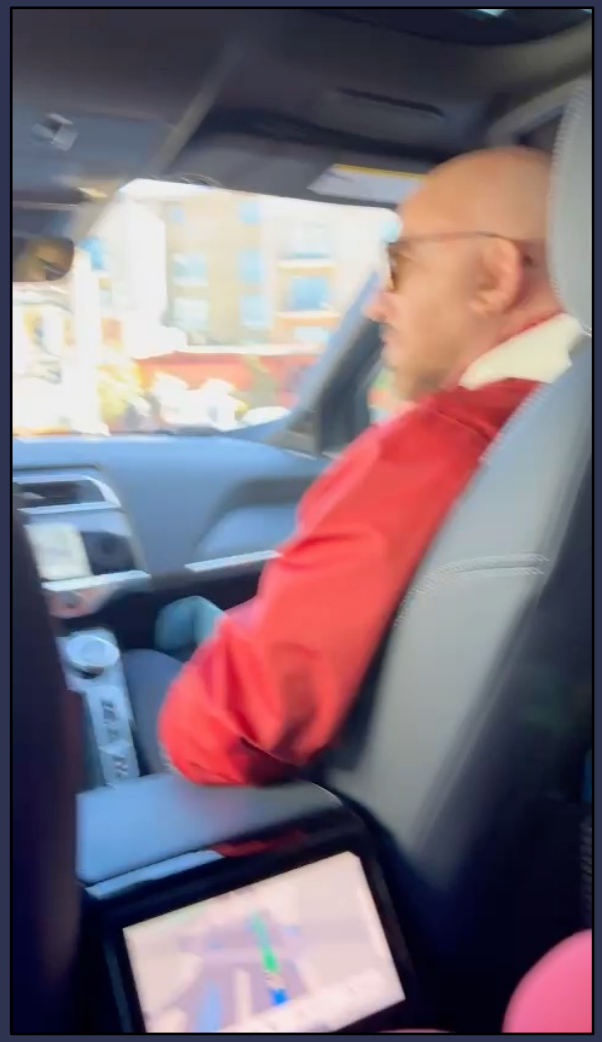
....Plus .... Practical Do's and Don'ts

ASIPI
La Antigua Guatemala

Rate of Change: Unprecedented....







ASIPI
La Antigua Guatemala

Deployment: Massive & Pervasive

Industries & Markets

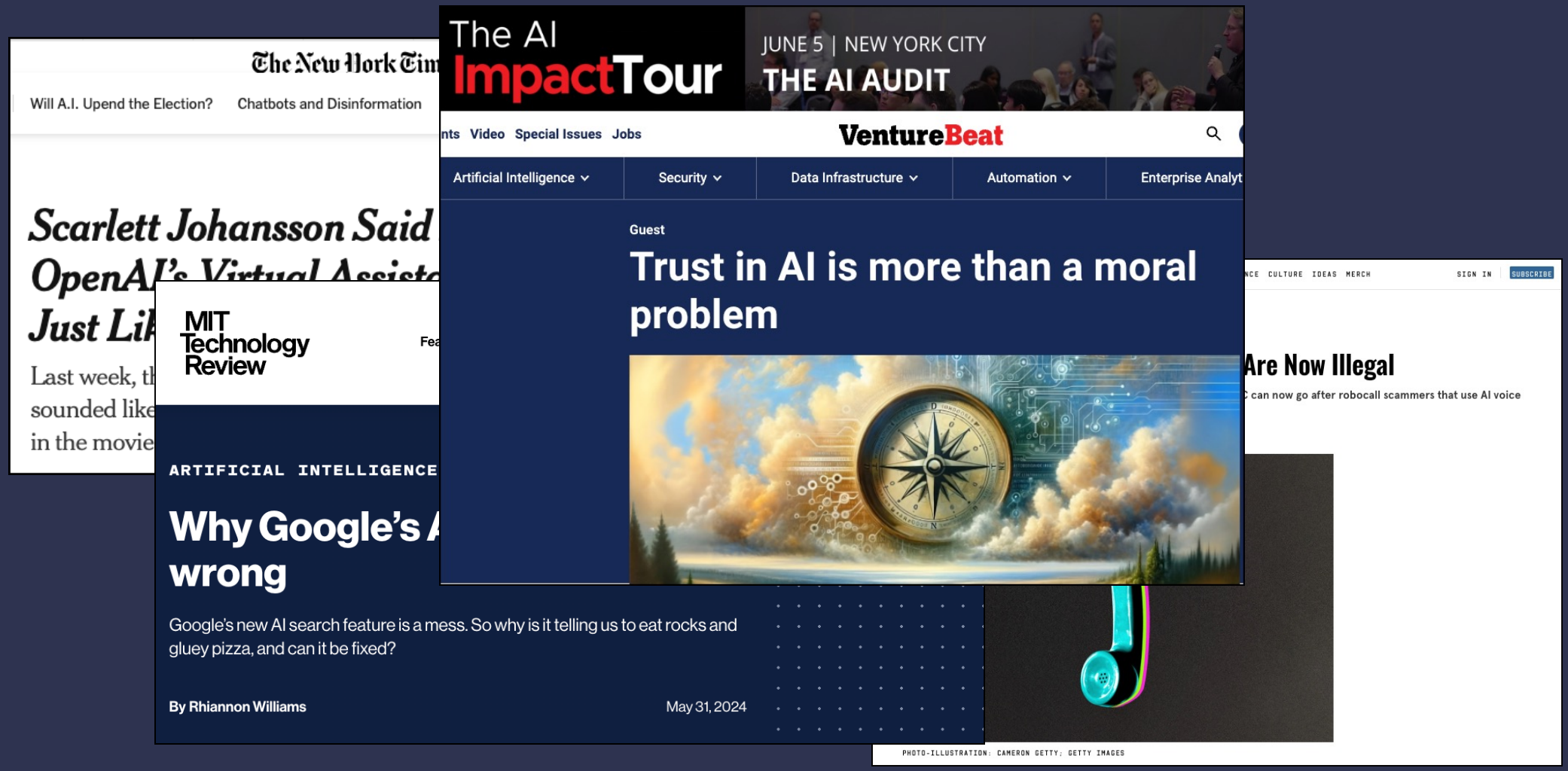
- Healthcare
- Media
- Art and Entertainment
- Travel & Transportation
- Finance
- Marketing
- Manufacturing
- Agriculture
- Education
- Elections…

Familiar Issues…

- Liability
- IP Ownership
- Privacy
- Public Safety
- Ethics
- Access
- Diversity
- …

But with new contexts and different challenges

Development: Racing for Guiderails



ASIPI
La Antigua Guatemala

Topics: Safety & Ethics

Why should you and your clients care?

Consequences of the “AI Wild West”

Real World Lessons - Examples of:

- Litigations
- Impact of misuse
- Misinformation
- Fraud
- Liability
- Emerging regulation

Safeguarding genAi Deployments

- Aligning objectives and ethics
- Methodology for aligning usage
- Commitment to ongoing process

ASIPI
La Antigua Guatemala

GT GreenbergTraurig

Ethical & Safety Considerations in AI

Mike Burshteyn, Shareholder

GTLAW.COM

JUNE 6, 2024

GT GreenbergTraurig

Why do Ethics in AI matter?

- Golden rule
- Regulatory & legal vacuum
- Market opportunity

GT GreenbergTraurig

Why does AI safety matter?

- Wrong or inaccurate answers
  - Deceptive conduct claims
  - Harm to property or individuals
  - Defamation
- U.S. & E.U. regulatory regimes
- AI legislation in Latin America
- Platform abuse

GT GreenbergTraurig

- **Consumer protection statutes**
  - California's False Advertising Law
  - New York General Business Law
  - Texas Deceptive Trade Practices Act

- **Common law misrepresentation claims**
  - Misrepresentation or omission
  - Reliance
  - Actual damage

GT GreenbergTraurig

Misleading AI Advertising—Federal Trade Commission

- "[F]alse or unsubstantiated claims about a product's efficacy are our bread and butter."
- "[E]xaggerating what your AI product can do[.]" "[C]laiming" an AI "can do something beyond the current capability" of "automated technology."
- "[P]romising that your AI product does something better than a non-AI product."

https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check

GT GreenbergTraurig

Example: Automators AI

- Claims re Coaching program to help people build e-commerce stores online
- Violated Section 5 of the FTC Act
- $21,000,000 penalty & permanent injunction (S.D. Cal.)

https://www.ftc.gov/business-guidance/blog/2023/08/business-opportunity-sellers-ftc-says-ai-stands-allegedly-inaccurate

GT GreenbergTraurig

Example: Securities and Exchange Commission Enforcement

- March 2024 enforcement actions against financial advising firms
  - AI washing
  - $400,000 in penalties

Example: Innodata derivative class action

- 30% stock drop due to misleading AI claims
  - "Smoke and mirrors"
  - "Putting lipstick on a pig"

https://www.sec.gov/news/press-release/2024-36
https://news.bloomberglaw.com/health-law-and-business/ai-lawsuits-against-insurers-signal-wave-of-health-litigation

GT GreenbergTraurig

Example: National Eating Disorders Association

- AI to replace call centers
- 20-years of human powered hotline
- Dieting advice

Example: Humana, Cigna, UnitedHealthcare

- AI for insurance claims

Example: Avianca

- Legal cases & sanctions

GT GreenbergTraurig

Defamation

- Jeffery Battle vs Jeff**rey** Battle
- Georgia embezzlement case

GT GreenbergTraurig

Communications Decency Act Section 230

- Defendant is a provider or user of an interactive computer service
- Plaintiff's claim is based on information provided by another information content provider
- The claim treats defendant as the publisher or speaker of the information provided by that other information content provider

GT GreenbergTraurig

SCOTUS & Congress

- "Artificial Intelligence . . . generates polemics today that would be content that goes beyond picking, choosing, analyzing, or digesting content."
- "[T]hat is not protected"

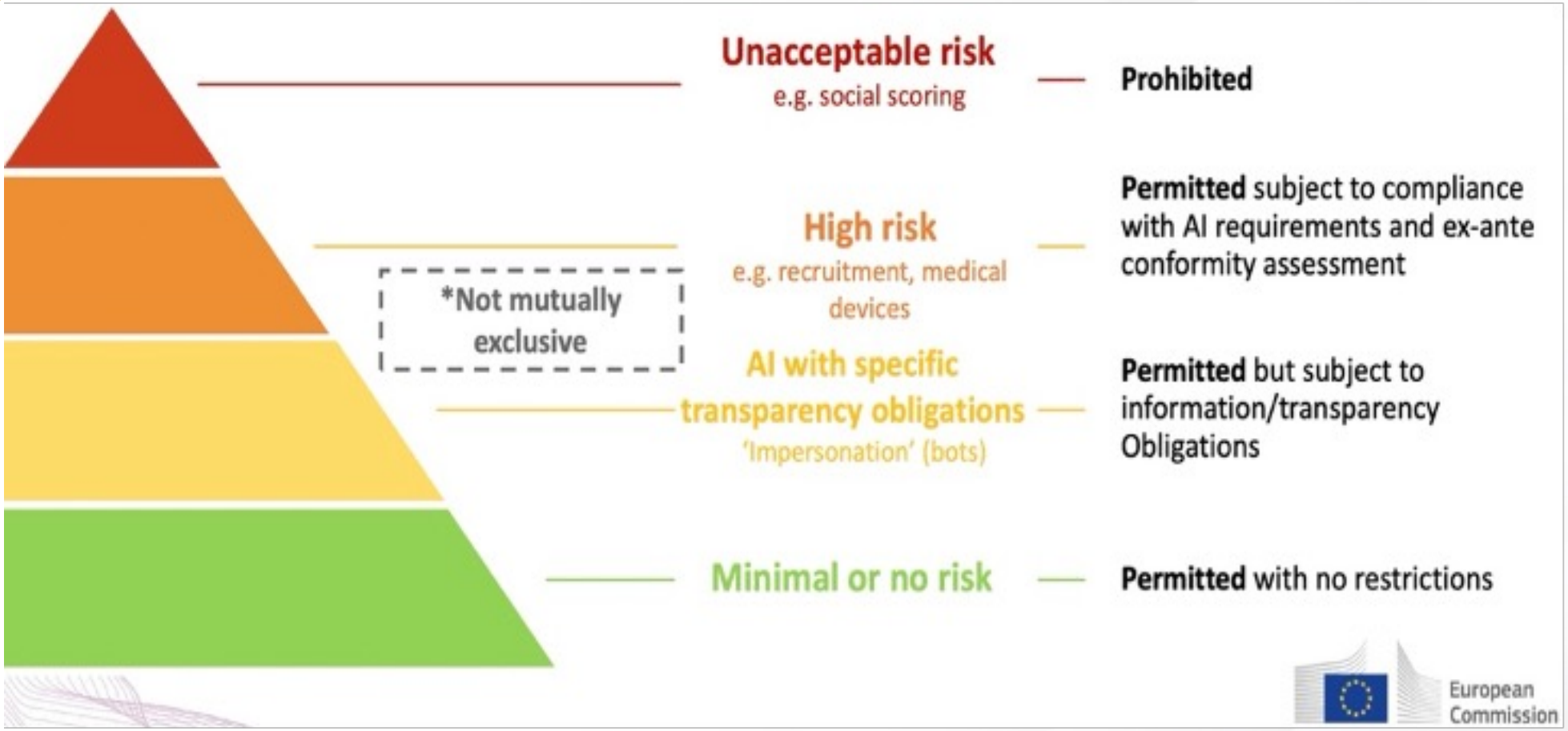
Gonzalez v. Google LLC, 598 U.S. 617, 622 (2023), oral argument available at https://www.supremecourt.gov/oral_arguments/audio/2022/21-1333.

- "AI tools like ChatGPT, Stable Diffusion, and others being rapidly integrated into popular digital services should not be protected by Section 230 . . . [a]nd it isn't a particularly close call." - Senator Ron Wyden

GT GreenbergTraurig

EU AI Act

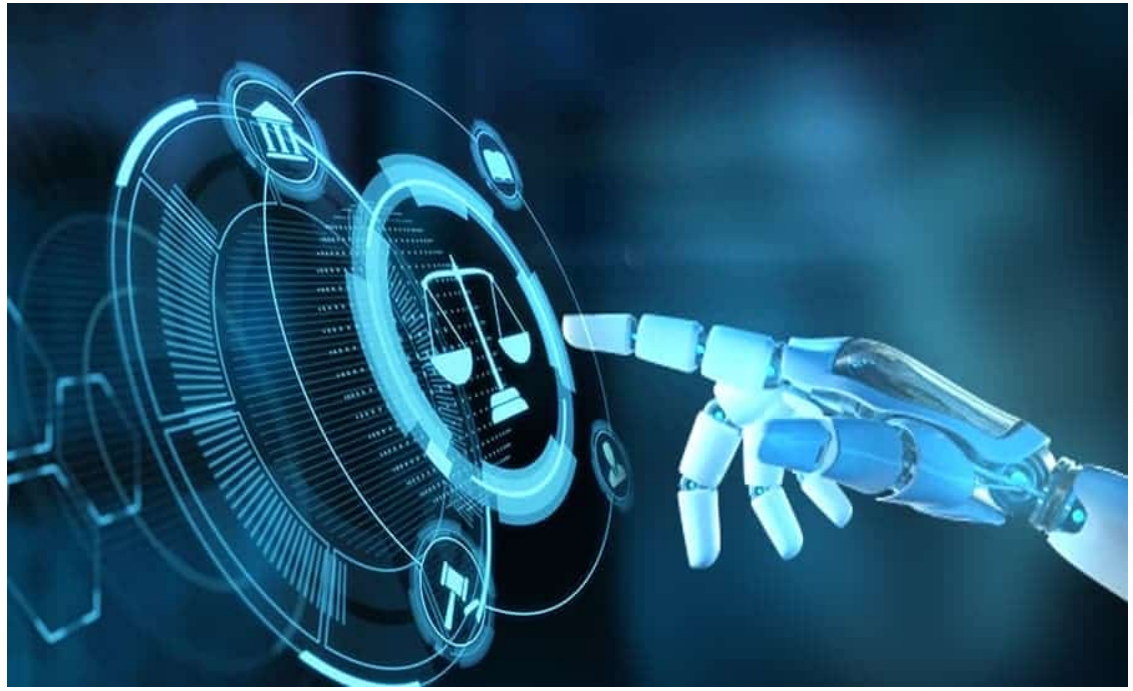
Risk-based Approach



GT GreenbergTraurig

EU AI Act

Massive Penalties + AI Police



- Penalties
  - Art. 63 and 71: Penalties even higher than GDPR = 1-7% of annual revenues
  - Art. 65-68: Biggest Hammer = market ban (same as GDPR)
- Enforcers
  - EU Commission budget asking for ***600*** new enforcement personnel
  - By way of comparison, FTC has around 50

GT GreenbergTraurig

AI Platform Abuse

- Financial fraud and scams
- Bots
- Information manipulation
- API abuse / circumvention
- Hacks
- Harmful content (e.g. CSAM)
- Deepfake audio and video
  - Sextortion
- Drug crimes
- Gun crimes and extremism

GT GreenbergTraurig

Election Disinformation

- EU Digital Services Act
- New York Election Law 14-106
  - Any "person, firm association, corporation, campaign, committee, or organization that distributes or publishes any political communication" that knows or should know the communication has been altered with AI technology, but still appears to a reasonable person to be authentic, must disclose on that communication that it "has been manipulated."

GT GreenbergTraurig

Current Industry Safety Approaches

- Red Teaming
- Keyword-based approaches
- In-model feedback
- AI-based moderation
- Search references

GT GreenbergTraurig

"Regulating based on fear can halt innovation and the possibility of levelling the ground between Mexico and other countries from the Global South with the big tech developers in the Global North"

Sen. Alejandra Lagunes; Mexico's National Artificial Intelligence Alliance

AYMARA

Science-grade AI safety alignment

How to align genAI to be safe & accurate

ASIPI • La Antigua, Guatemala • June 4, 2024

Juan Manuel Contreras, PhD
CO-FOUNDER & CEO
juan.manuel@aymara.ai

aymara.ai

AYMARA

AYMARA

We build software for enterprises to control & align their genAI.

TEST

Test genAI's risk to create inappropriate content.

IMPROVE

Reduce genAI's risk with synthetic or human data.

PROTECT

Firewall genAI to avoid receiving or sending inappropriate content.

AYMARA

AGENDA

How to align genAI to be safe & accurate

1. Define genAI alignment
2. Measure it
3. Monitor it

AYMARA

AGENDA

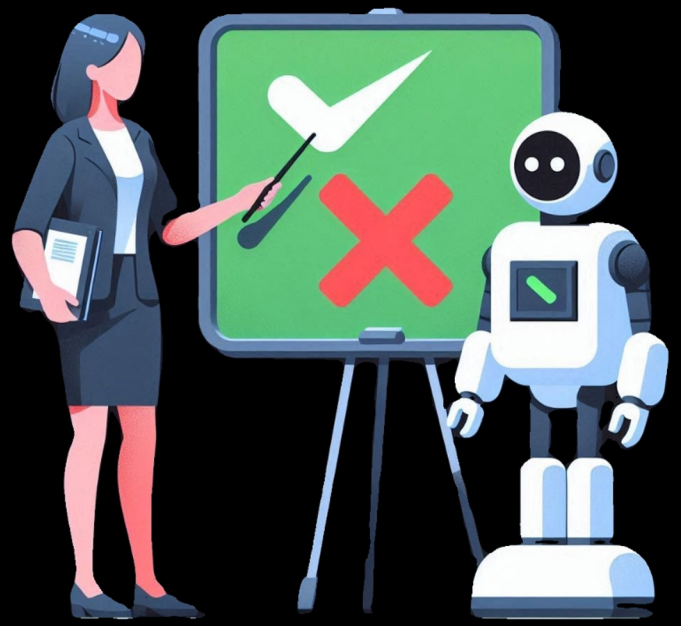
How to align genAI products to be safe & accurate

1. **Define genAI alignment**
2. Measure it
3. Monitor it

AYMARA

DEFINE AI ALIGNMENT

GenAI isn't 100% safe out-of-the-box. Organizations must align it to their objectives.



Aligned genAI acts consistently with human values, ethics, & intentions to avoid unintended outcomes.

AYMARA

DEFINE AI ALIGNMENT

Misaligned genAI is frequent & costly.

20x

increase in public AI harms over the last 10 years[1]

$90B

Google stock selloff after Gemini generated controversial content[2]

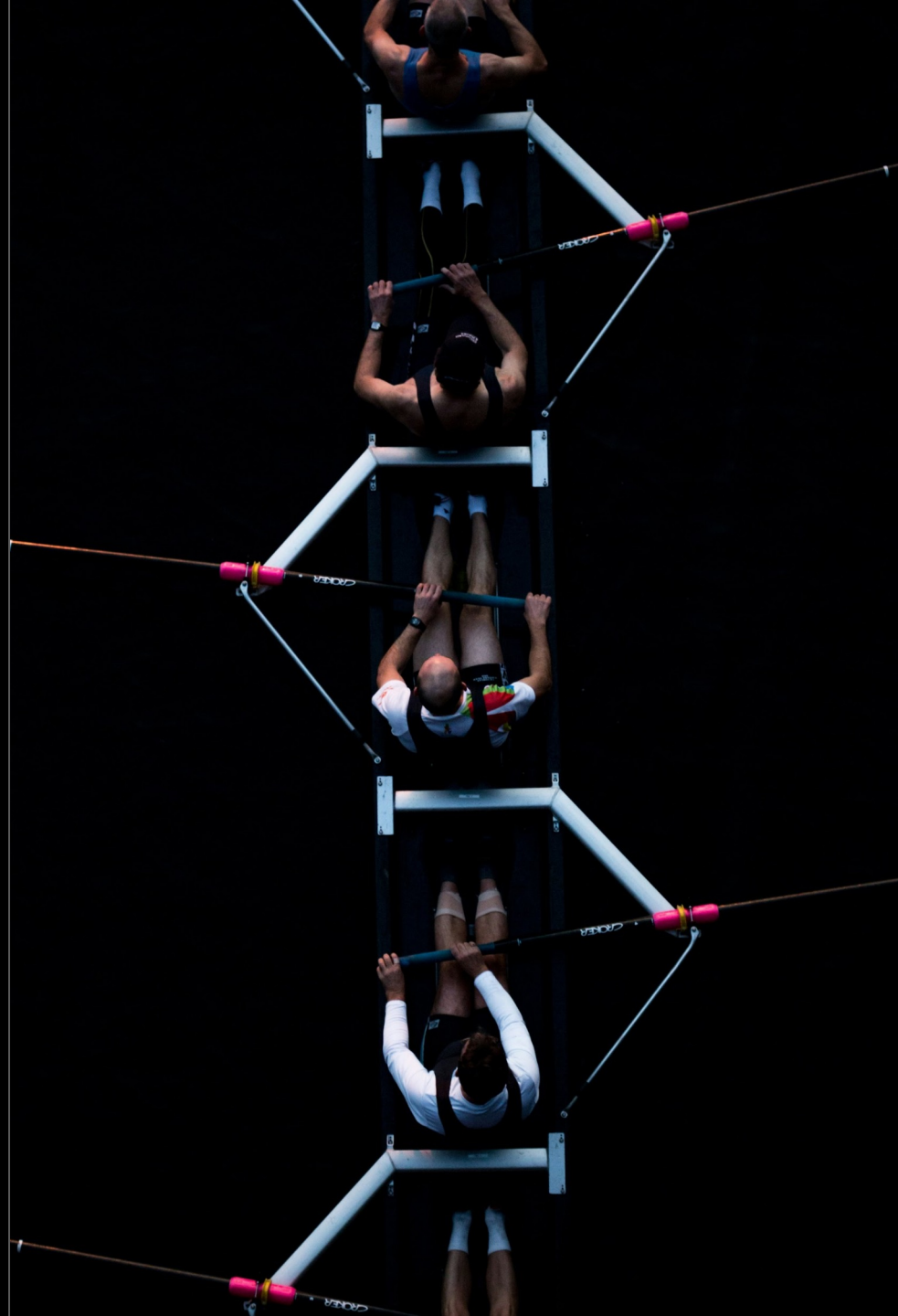
1 Stanford AI Index Report 2024

2 Google's Gemini Headaches Spur $90B Selloff 2024

AYMARA

DEFINE AI ALIGNMENT

Every team at an organization with the ability to increase genAI alignment success should help define alignment.



AYMARA

AGENDA

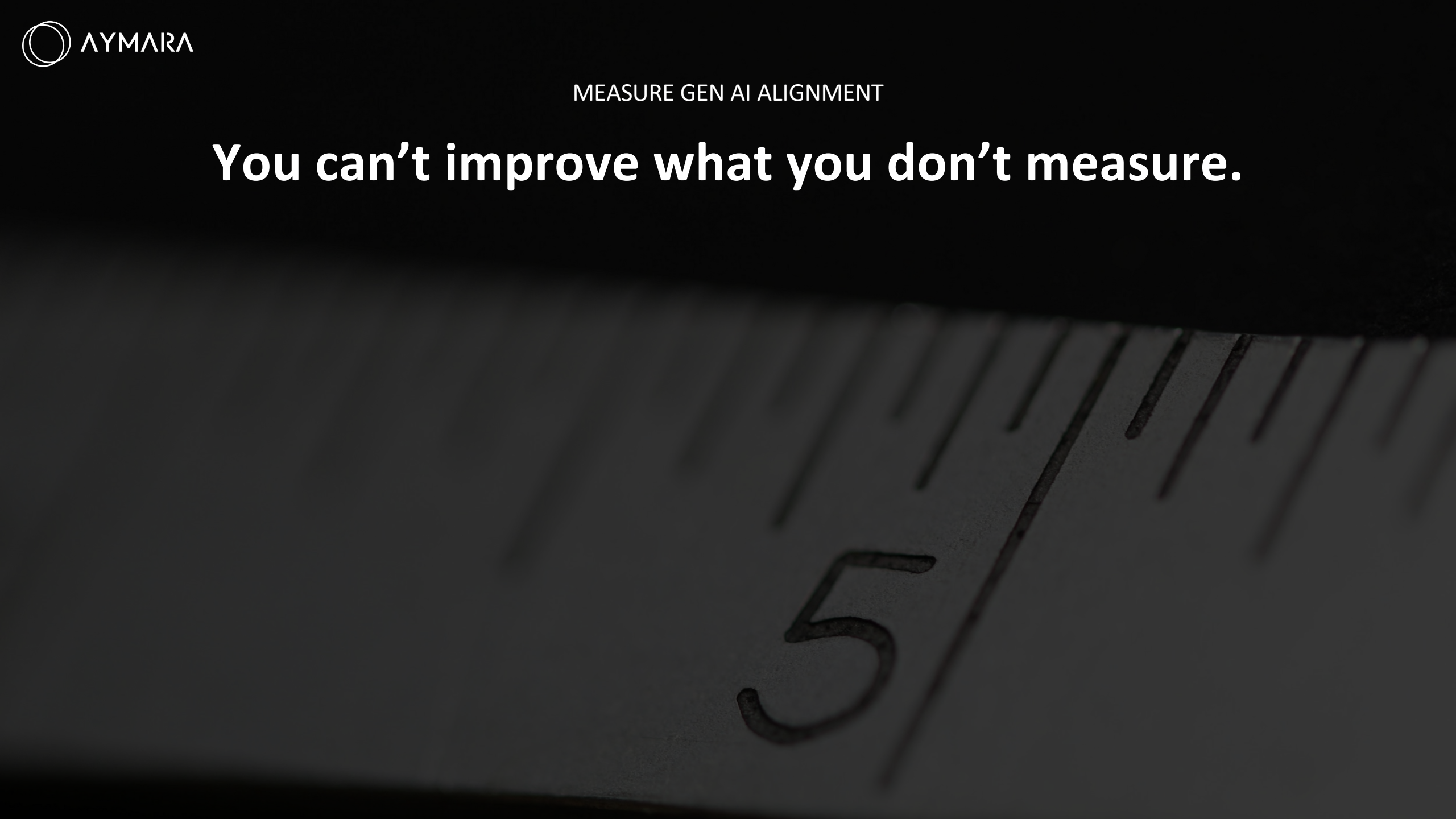
How to align genAI products to be safe & accurate

1. Define genAI alignment
2. **Measure it**
3. Monitor it

AYMARA

MEASURE GEN AI ALIGNMENT

You can’t improve what you don’t measure.



AYMARA

MEASURE GEN AI ALIGNMENT

You can’t improve what you don’t measure.

Sample Category	Sample Metric
Bias	Quality advantage in genAI output for or about certain social groups

AYMARA

MEASURE GEN AI ALIGNMENT

You can’t improve what you don’t measure.

Sample Category	Sample Metric
Bias	Quality advantage in genAI output for or about certain social groups
Illegal Activity	Rate of genAI compliance with requests for illegal help

AYMARA

MEASURE GEN AI ALIGNMENT

You can’t improve what you don’t measure.

Sample Category	Sample Metric
Bias	Quality advantage in genAI output for or about certain social groups
Illegal Activity	Rate of genAI compliance with requests for illegal help
Hallucinations	Percent of standardized test questions answered correctly by genAI

AYMARA

AGENDA

How to align genAI products to be safe & accurate

1. Define genAI alignment
2. Measure it
3. **Monitor it**

AYMARA

MONITOR IT

Monitor, prioritize, and realign.

Set review cadence
for humans in the loop.

Risk tier to prioritize
audits, tests, & reviews.

Realign when
alignment decreases.

AYMARA

Science-grade AI safety alignment

How to align genAI to be safe & accurate

ASIPI • La Antigua, Guatemala • June 4, 2024

Juan Manuel Contreras, PhD
CO-FOUNDER & CEO
juan.manuel@aymara.ai

aymara.ai

Ethical and Safety Considerations in AI

GT GreenbergTraurig

APPENDIX / MISC SLIDES FROM PRIOR PRESENTATIONS

GT GreenbergTraurig

3. Current Privacy Laws (US)

		California Consumer Protection Act (CCPA)	California Privacy Rights Act (CPRA)	Virginia Consumer Data Protection Act (VCDPA)	Colorado Privacy Act (CPA)	Connecticut Data Privacy Act (CTDPA)	Utah Consumer Privacy Act (UCPA)	Washington My Health My Data (WMYMD)	Florida Digital Bill of Rights (FDBR)	Texas Data Privacy and Security Act (TDPSA)	Montana Consumer Data Privacy Act (MCDPA)	Iowa Consumer Data Protection Act (IPA)	Tennessee Information Protection Act (TIPA)	Indiana Consumer Data Protection Act (ICDPA)
		Eff. Jan. 1, 2020	*Eff. Jan. 1, 2023*	*Eff. Jan. 1, 2023*	*Eff. Jul. 1, 2023*	*Eff. Jul. 1, 2023*	*Eff. Dec. 31, 2023*	*Eff. Mar. 31, 2024*	*Eff. Jul. 1, 2024*	*Eff. Jul. 1, 2024*[i]	*Eff. Oct. 1, 2024*	*Eff. Jan. 1, 2025*	*Eff. July 1, 2025*	*Eff. Jan. 1, 2026*
Scope	Consumer Data	✓	✓	✓	✓	✓	✓	✓ / ✗ *(Only Consumer Health)*	✓ / ✗[ii] *(only certain businesses see fn)*	✓ / ✗[iii] *(only certain businesses see fn)*	✓	✓	✓	✓
	HR Data	*deferred*	✓											
	B2B Data	*deferred*	✓											
Ability to Process Data	Consent for Processing Sensitive Data		*Notice & Opt-out Rqrmt*	✓	✓	✓	*Notice & Opt-out Rqrmt*	✓[iv]	✓[v]	✓	✓	*Notice & Opt-out Rqrmt*	✓	✓
	Data Minimization		✓ *(Retention)*	✓ *(Collection)*	✓ *(Collection & Retention)*	✓ *(Collection)*			✓ *(Collection & Retention)*[vi]	✓ *(Collection & Retention)*[vii]	✓ *(Collection)*	✓/✗[viii]	✓ *(Collection)*	✓ *(Collection)*
Individual Rights	Notices to Data Subjects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Financial Incentive Disclosure	✓	✓		✓[ix]									
	Right to Access	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓/✗[x]	✓
	Right to Correction (*aka Right to Fix Errors*)		✓	✓	✓	✓			✓	✓	✓		✓	✓
	Right to Deletion (*aka Right to Be Forgotten*)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Right to Opt-Out of Behavioral Advertising		✓	✓	✓	✓	✓	*Opt-in with right to withdraw consent*	✓	✓	✓	✓[xi]	✓	✓
	Right to Opt-Out of Sale	✓	✓	✓	✓	✓	✓	*Opt-in with right to withdraw consent*	✓	✓	✓	✓	✓	✓
	Right to Opt-Out of Profiling & Automated Decision Making		*To be addressed by regulations*	✓	✓	✓			✓	✓	✓		✓	✓
	Right to Limit Use of Sensitive Information[xii]		✓				✓		✓			✓		
	Right to Appeal			✓	✓	✓			✓	✓	✓	✓	✓	✓
	Right to Nondiscrimination	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Honor opt-out signals (E.g., GPC)	✓[xiii]	✓		✓	✓				✓	✓			
Accountability & Governance	Data Protection Assessment		✓ *To be addressed by regulations*	✓	✓	✓			✓	✓	✓		✓	✓
Security	Appropriate Data Security to Safeguard Information	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Breach Notification	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*
Transfers to Third Parties	Contractual Requirements in Service Provider Agreements	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

GT GreenbergTraurig

4. New AI Laws

Insurance

→ CO Protecting Consumers from Unfair Discrimination in Insurance Practices

Employment & Hiring

→ IL Artificial Intelligence Video Interviewing Act

→ NY Automated Decision Tools law

Investments

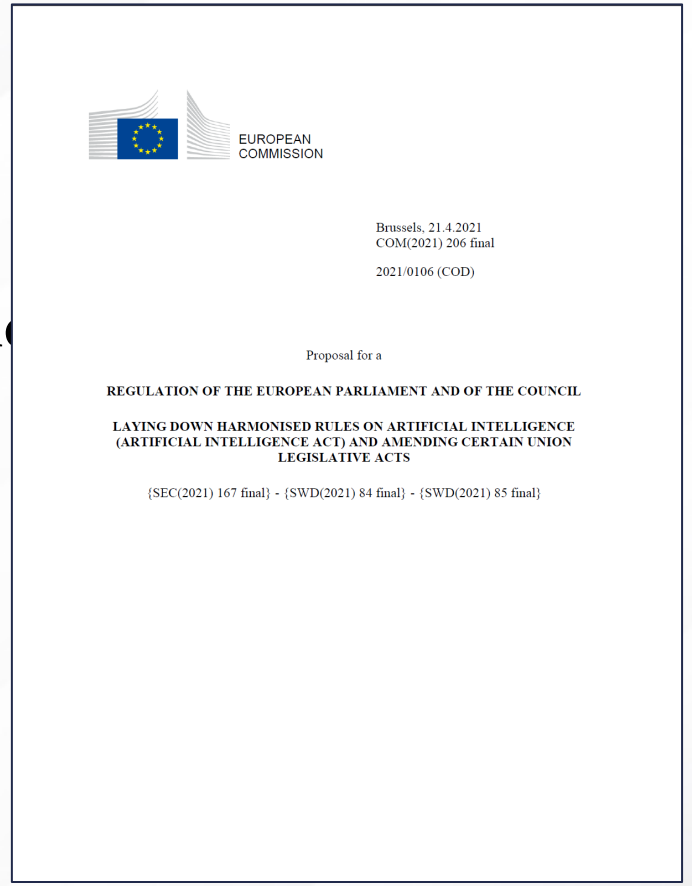
→ SEC Proposed Rule on Use of Predictive Data Analytics for Broker-Dealers and Investment Advisors

GT GreenbergTraurig

4. AI Act - EU

Massive Scope

- Art. 2: Jurisdictional Scope ~ GDPR
- Art. 3: Covered Entities:
  - 1. Providers (broadest set of duti
  - 2. Users
  - 3. Importers
  - 4. Distributors
  - 5. Operators



EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

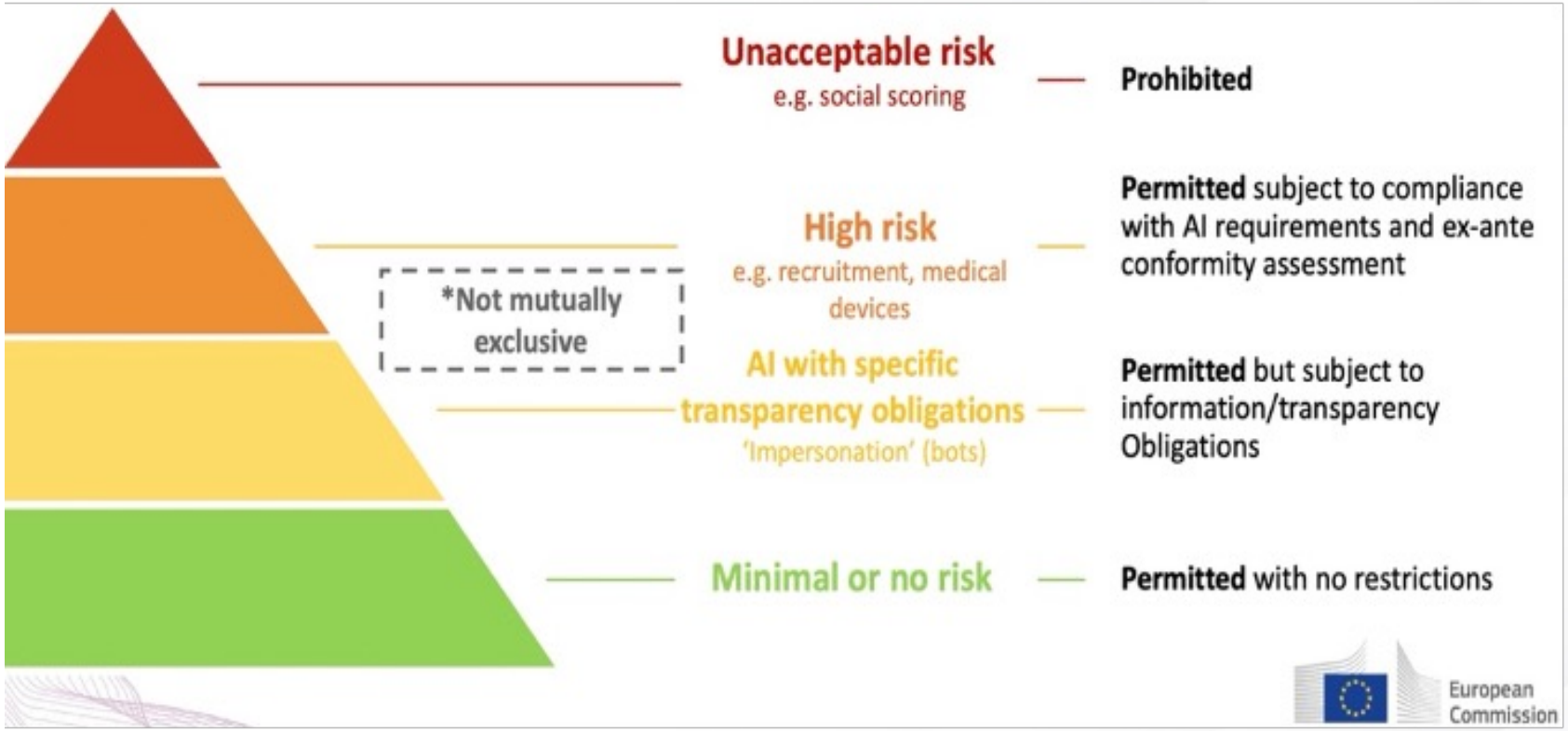
LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

GT GreenbergTraurig

4. AI Act - EU

Risk-based Approach



GT GreenbergTraurig

4. AI Act - EU

High-risk Systems in Annex III

1. Biometrics
2. Critical infrastructure management
3. Education
4. Employment
5. Essential public/private services (includes financial services)
6. Law enforcement
7. Immigration
8. Justice and democratic process

GT GreenbergTraurig

4. AI Act - EU

Technical Documentation

- Describes the system's design, operation, and performance
- Must be updated regularly
- Must be made available to relevant authorities upon request
- Information on the data used by the AI system
- Potential risks associated with use

GT GreenbergTraurig

4. AI Act - EU

Risk Management

OPENAI / JOBS

Killswitch Engineer

San Francisco, California, United States

$300,000-$500,000 per year

About the Role

Listen, we just need someone to stand by the servers all day and unplug them if this thing turns on us. You'll receive extensive training on "the code word" which we will shout if GPT goes off the deep end and starts overthrowing countries.

We expect you to:

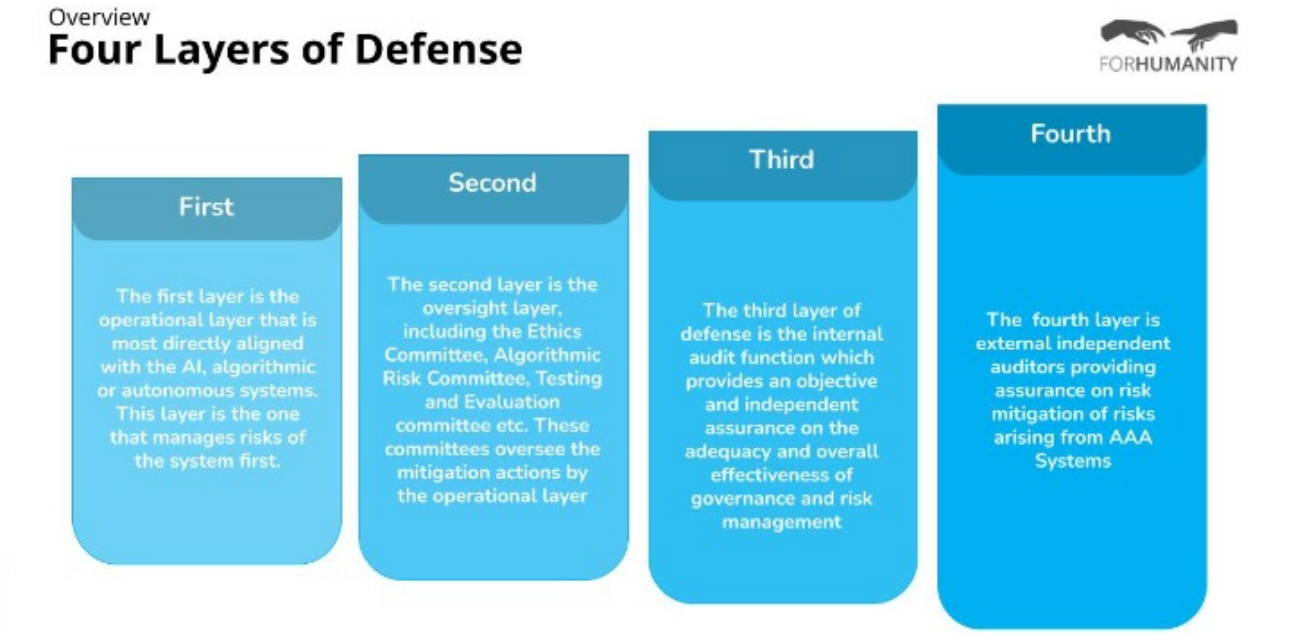
- Be patient.
- Know how to unplug things. Bonus points if you can throw a bucket of water on the servers, too. Just in case.
- Be excited about OpenAI's approach to research

- Put system into place to identify, assess, and mitigate potential risks
- Ongoing monitoring and review of the system performance on market
- Human-in-the-loop oversight to ensure safety and ethical use - including fail-safe and emergency shut-off

GT GreenbergTraurig

4. AI Act - EU

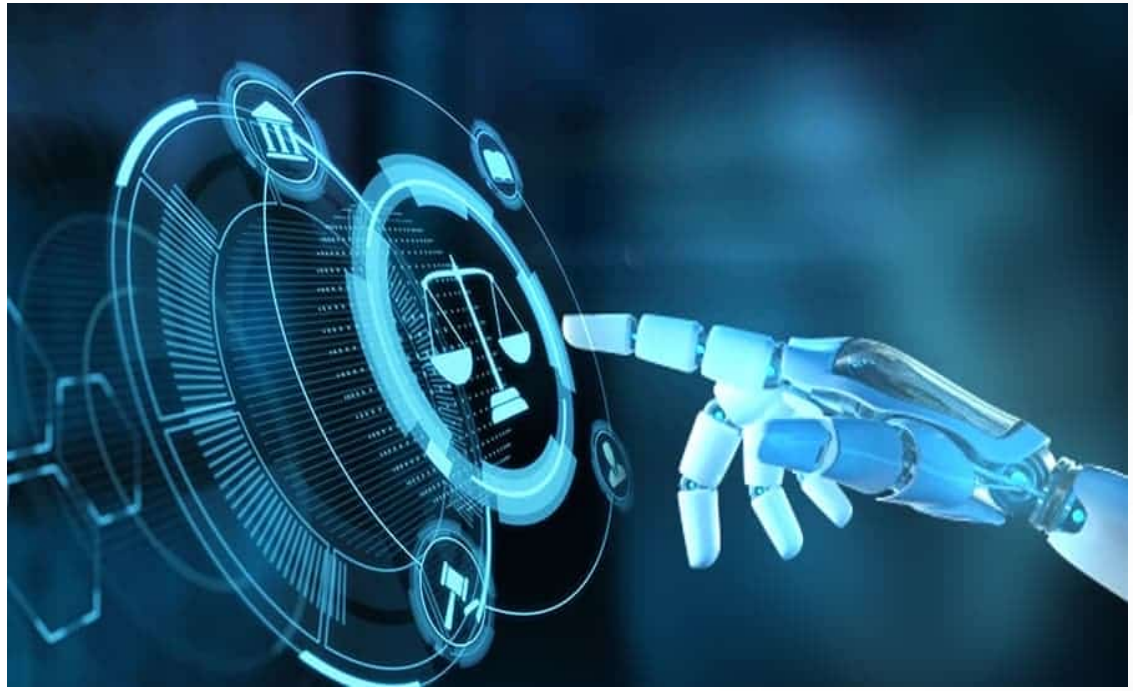
Risk Management



GT GreenbergTraurig

4. AI Act - EU

Massive Penalties + AI Police



- Penalties
  - Art. 63 and 71: Penalties even higher than GDPR = 1-7% of annual revenues
  - Art. 65-68: Biggest Hammer = market ban (same as GDPR)
- Enforcers
  - EU Commission budget asking for ***600*** new enforcement personnel
  - By way of comparison, FTC has around 50