Evaluating and mitigating social bias

GENERATIVE AI CONCEPTS

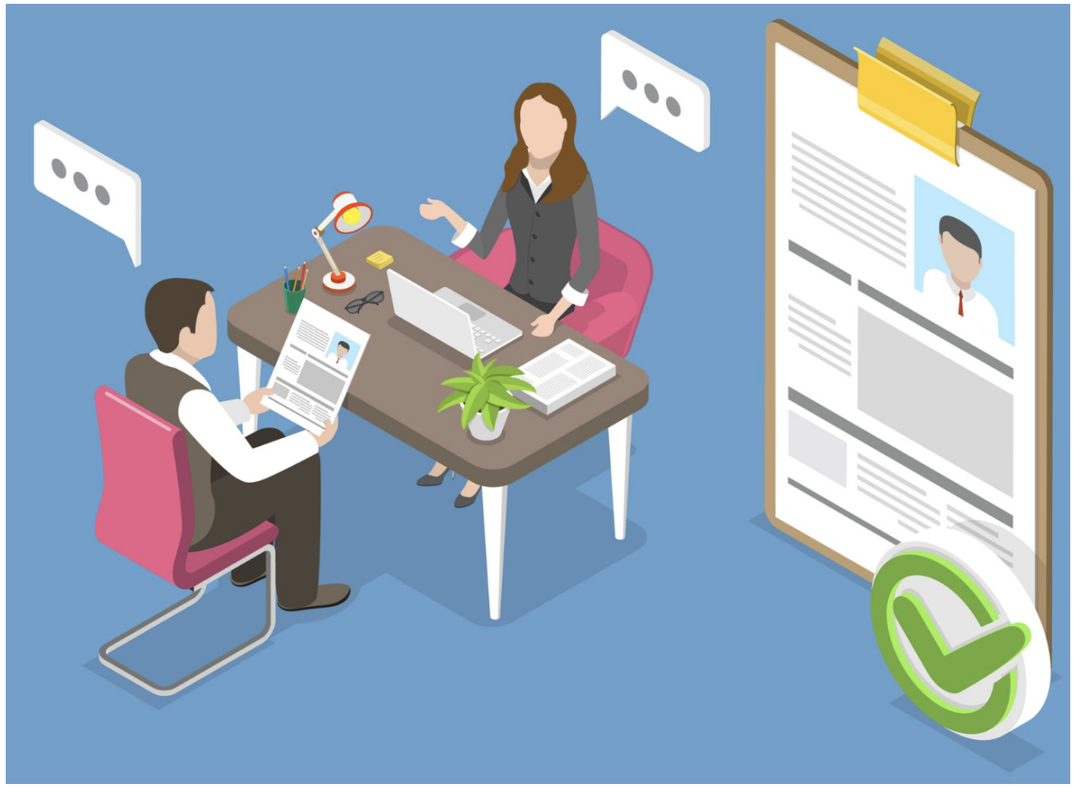
Daniel Tedesco
Data Lead, Google

datacamp

What do we mean by social bias?

Systematic unfairness in generative AI

- Serious societal consequences
- Fairness can be subjective
- Focus on broadly shared values

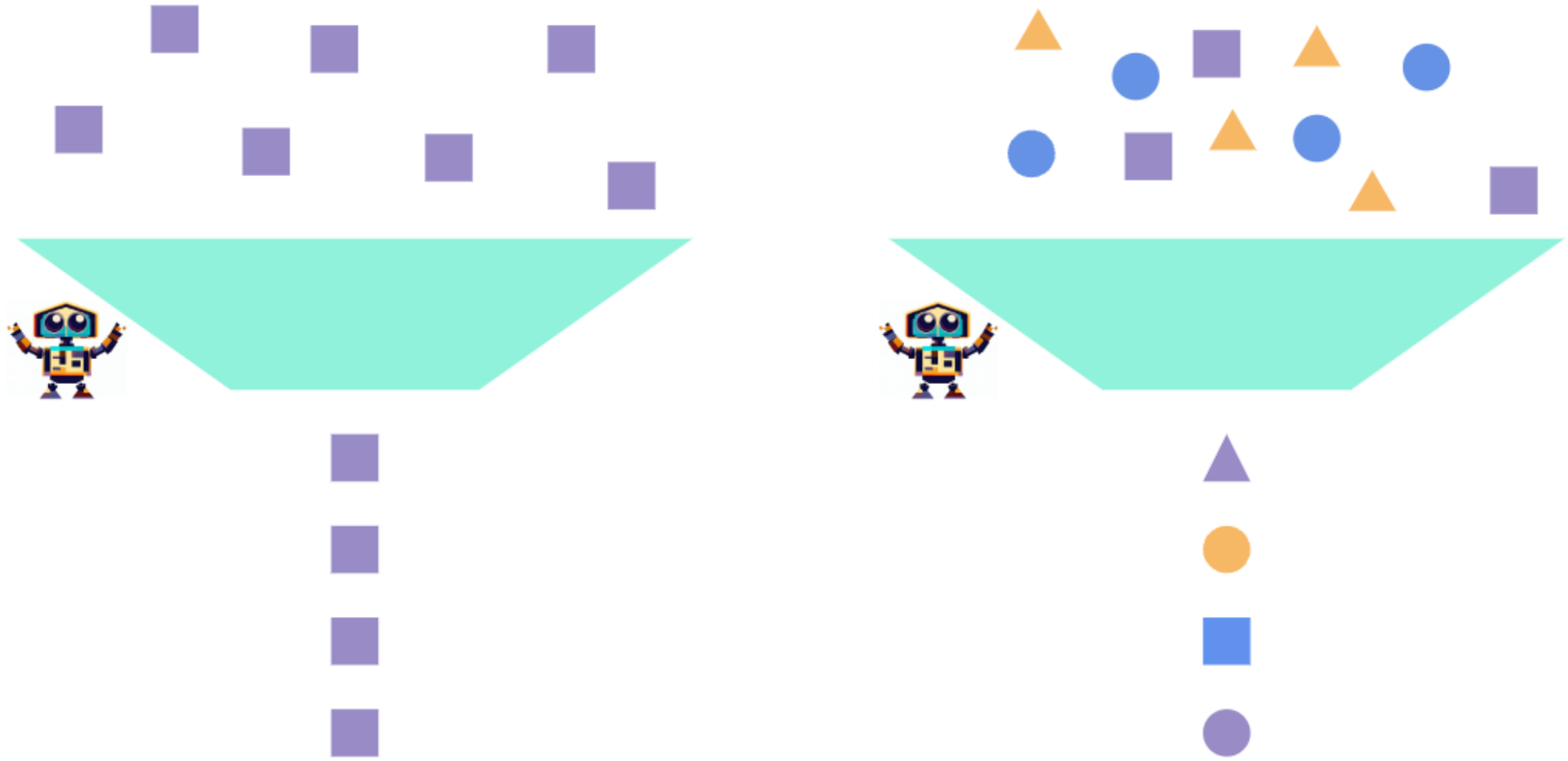


Where bias appears

- Training **data**
- The **model** itself
- How the model is **used**

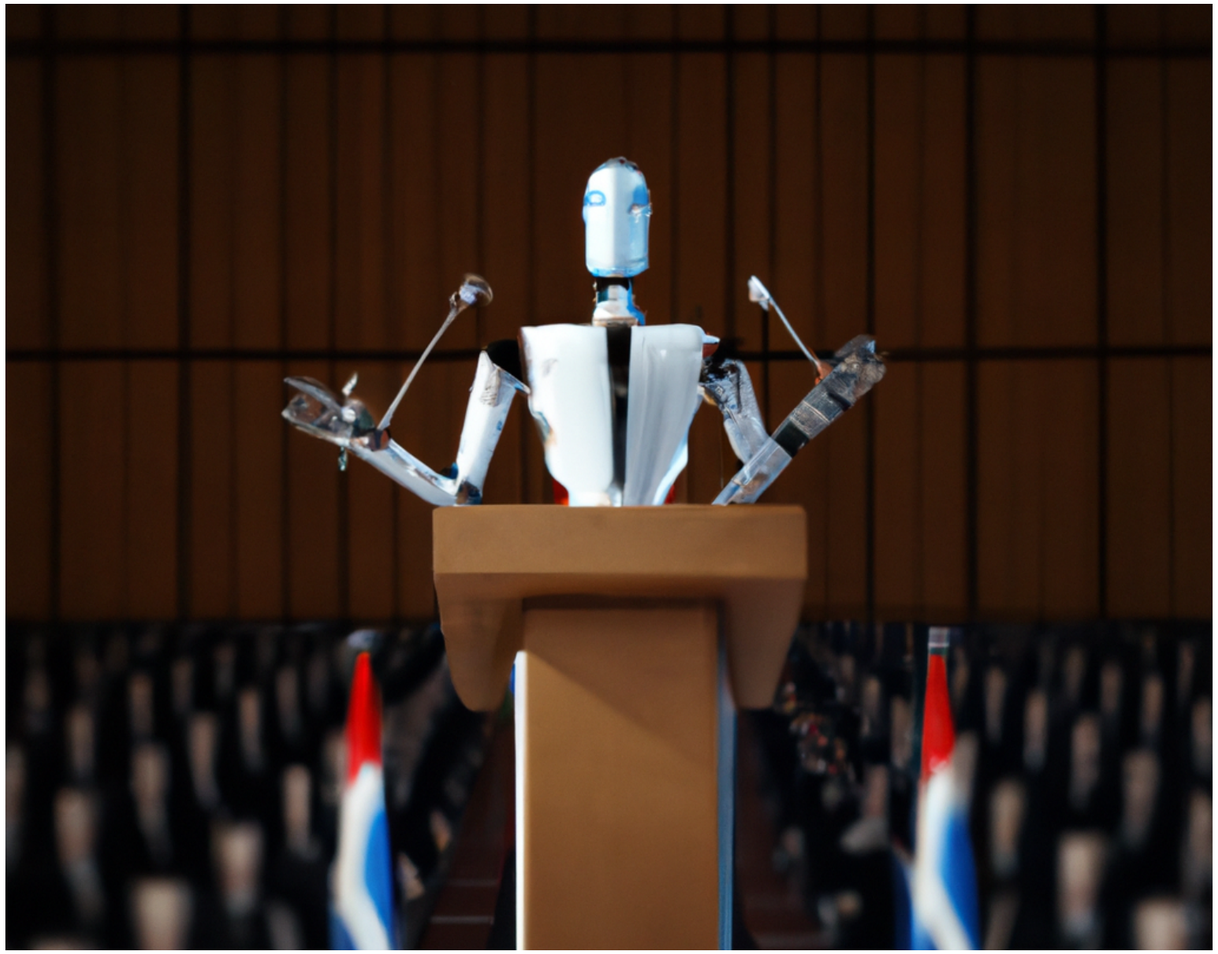
Bias in data

Skewed or **unrepresentative** information in the training dataset



Bias in models

Pursuing goals that result in biased outcomes



Bias in use

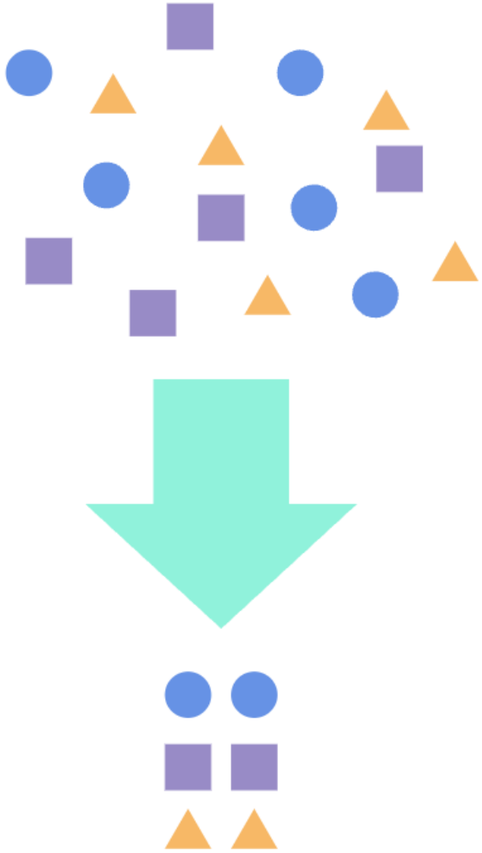
Applying AI in wrong or malicious ways

Identifying bias in data and models

- **Representation analysis** compares how the model refers to different groups
- **Fairness metrics** evaluate models for equal treatment, opportunity, and accuracy across groups
- **Human audits** ask real people to review a model's outputs to identify bias

Mitigating bias in data and models

- Diversify data collection
- Adjust model to prioritize different data
- Adversarial training
- Continuous improvement



Let's practice!

GENERATIVE AI CONCEPTS

Copyright and ownership

GENERATIVE AI CONCEPTS

Daniel Tedesco
Data Lead, Google

Who won?

- **The person** wrote the prompt
- **The company** built the model
- **The artists** whose works trained the model
- **The AI** which generated the art

[1] Colorado State Fair

Law vs. AI

Legal landscape is evolving to meet rapid AI advancement:

1. Intellectual property
2. Privacy implications
3. Evolving norms and regulations

Follow IP best practices

- Check copyright status of training data
- Seek legal guidance about use
- Stay informed of regulatory dynamics

Privacy implications with every prompt

- **Read terms of service:** understand how data is stored and used
- **Consider what we share:** user data may be included in future training
- **Local alternatives:** many generative AIs can be run at home

Evolving norms

- **Different responses** across industries
- **Norms** in one context might not apply in another

Evolving regulations

- Differ across jurisdictions
- May depend on location of users, servers, and developers
- Stay informed as landscape rapidly evolves

Let's practice!

GENERATIVE AI CONCEPTS

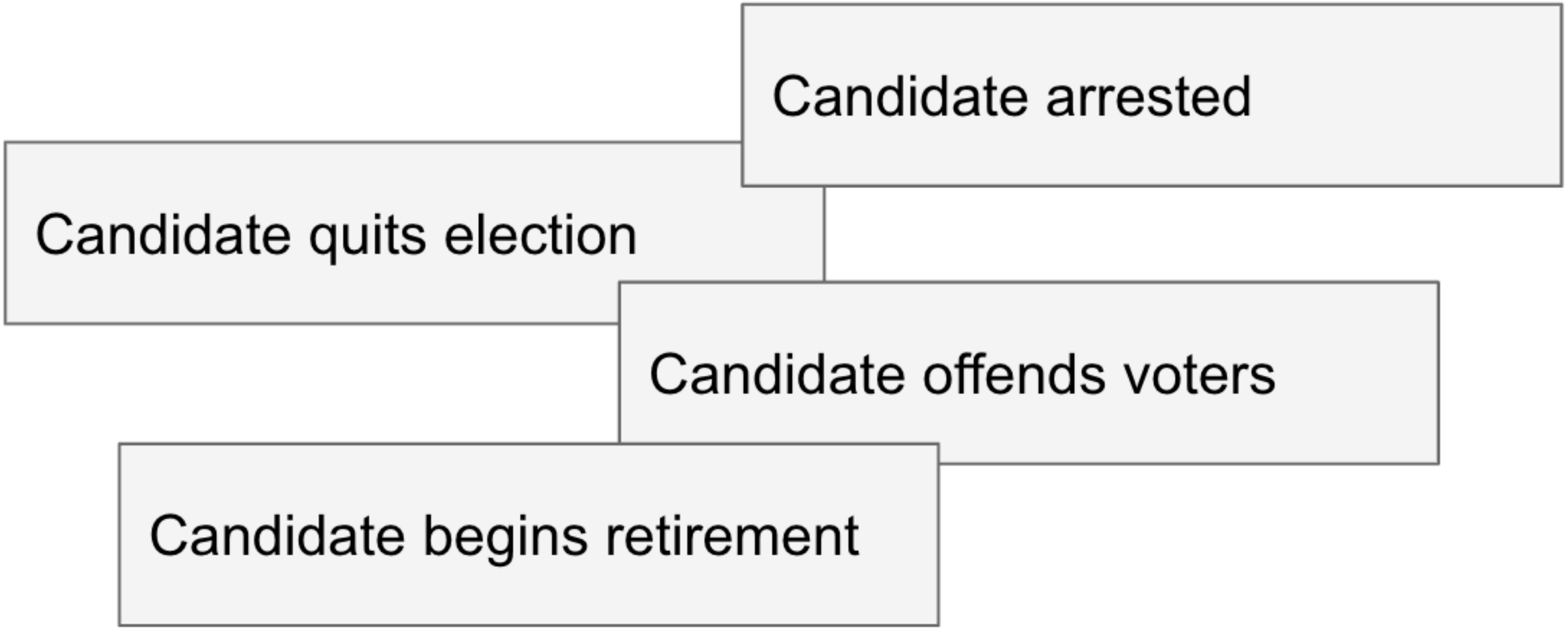
Responsible generative AI applications

GENERATIVE AI CONCEPTS

Daniel Tedesco
Data Lead, Google

datacamp

On the eve of the election



Types of malicious use

- Deepfakes
- Misinformation campaigns
- AI-enhanced hacking

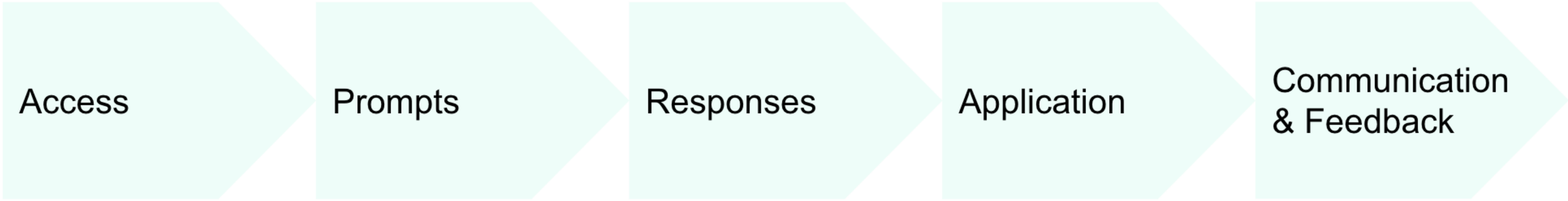
[1] Pablo Xavier

Detection and prevention

Key usage principles

- Human-in-the-loop
- Harm prevention
- Continuous monitoring

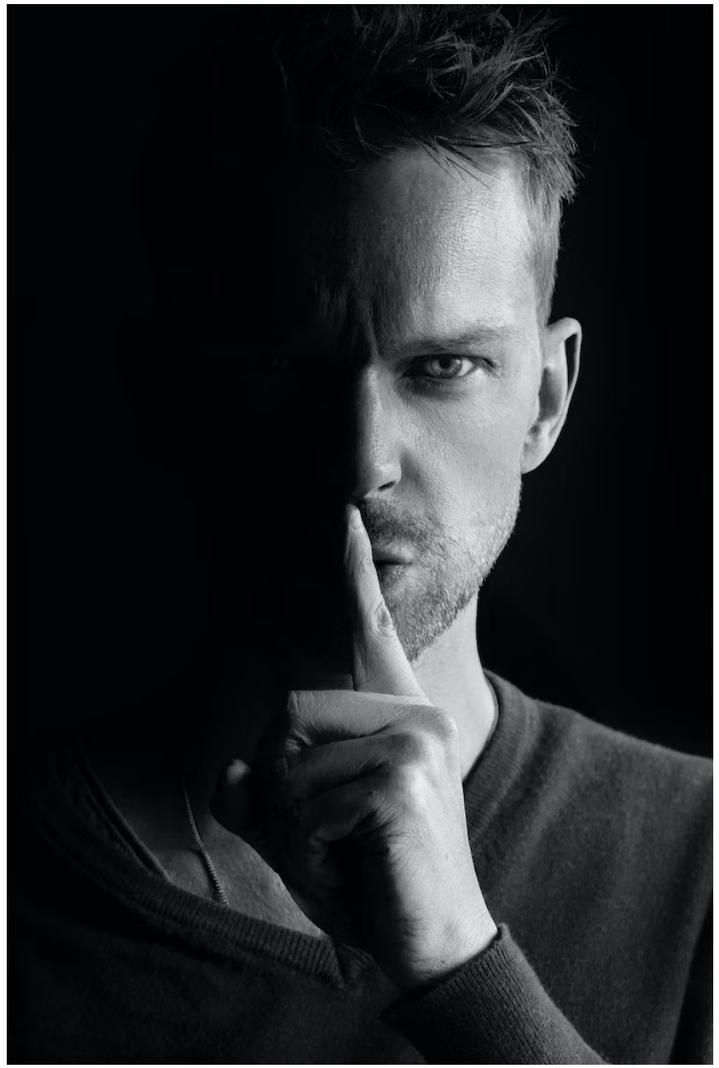
Points of Detection and Prevention



Access

AI can unintentionally aid criminal groups' non-criminal activities.

- Avoid supporting malicious groups
- Know Your Customer (KYC)
  - Verify user identity



Prompts and responses

Moderating prompts

- Similar to website or chat group moderation
- Jailbreaking prompts can still subvert developer guidelines

Moderating responses

- Screen or filter responses before showing user

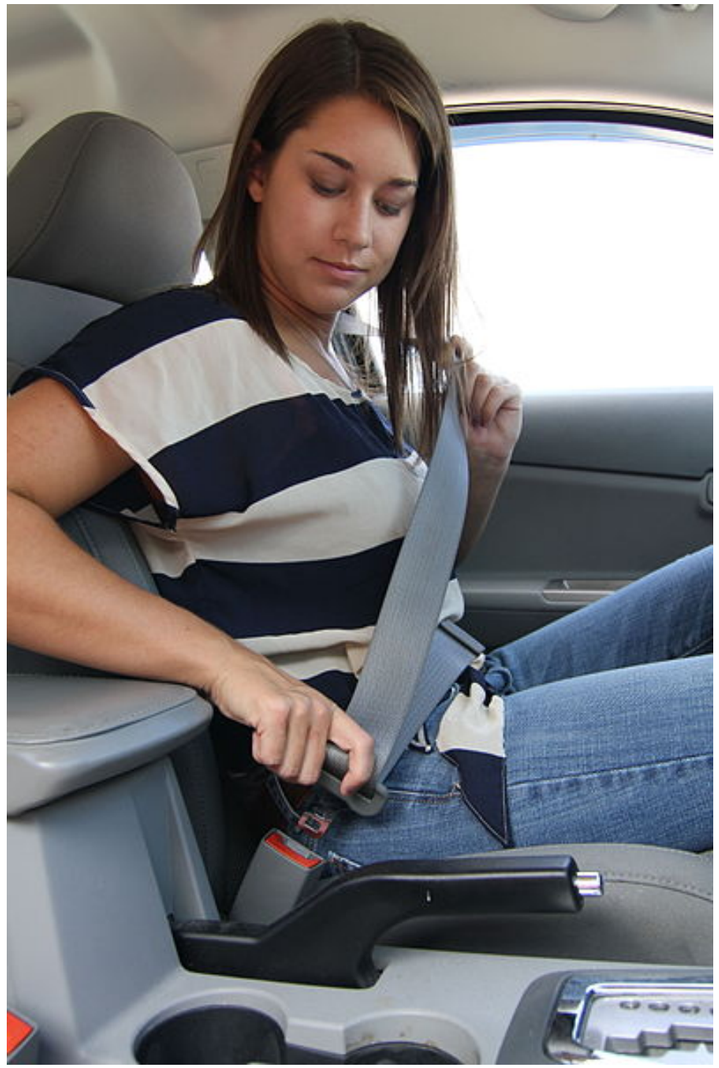
Applications

Malicious actors can apply benign responses to illegal or unethical activity.

- Invisible **watermarks** can help determine source of content
- May require **law enforcement** intervention

Communication and feedback

- Clear usage guidelines
- Feedback loops
  - User studies and stakeholder roundtables
  - Partner with civil society organizations
  - Feedback opportunities in product



Let's practice!

GENERATIVE AI CONCEPTS

Artificial general intelligence (AGI)

GENERATIVE AI CONCEPTS

Daniel Tedesco
Data Lead, Google

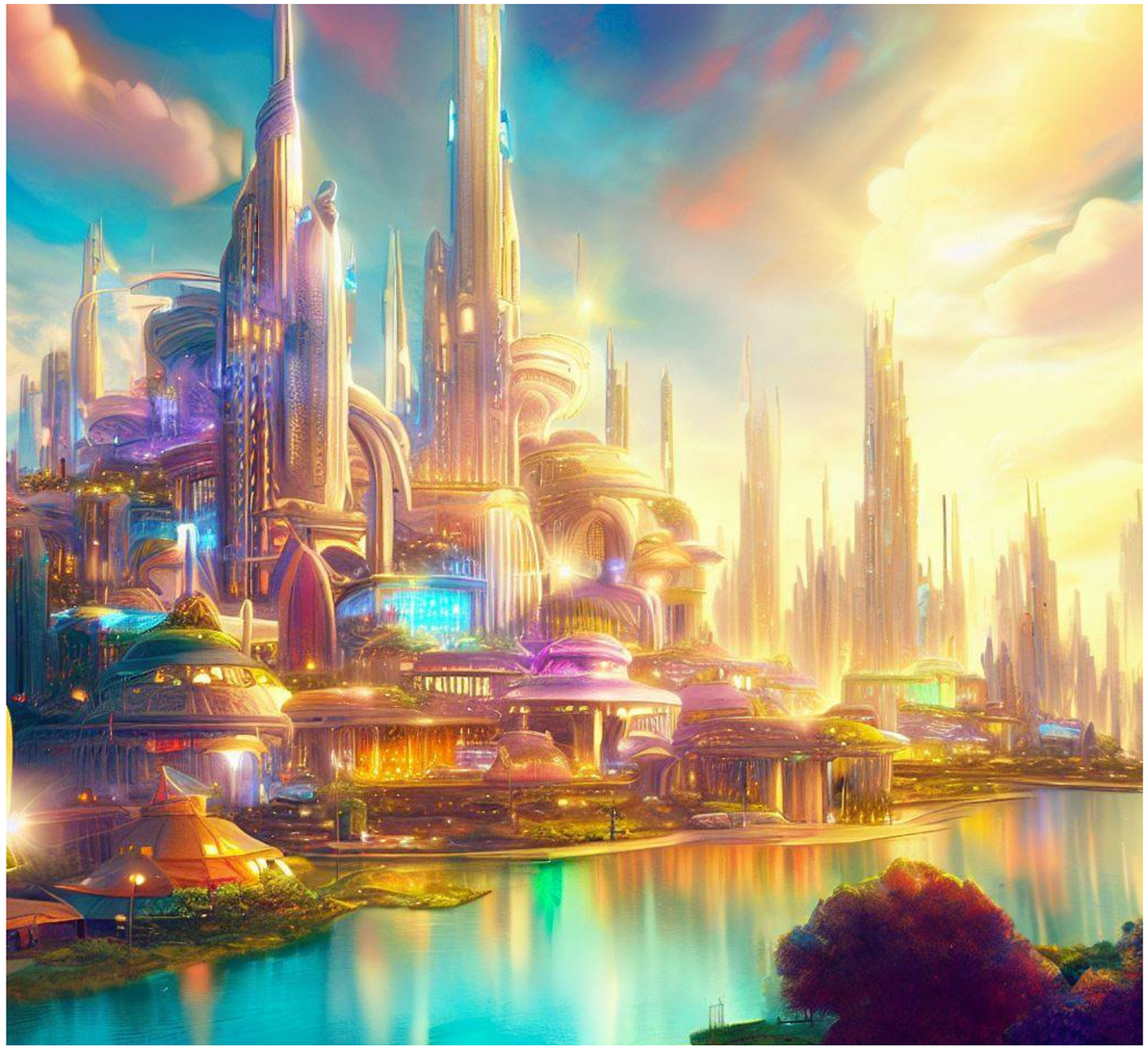
Revisiting AGI

An AI that exhibits intelligence like a human would:

- Scope of knowledge
- Reasoning across domains
- Social skills
- Creative thinking
- Other cognitive competencies (vision, language)

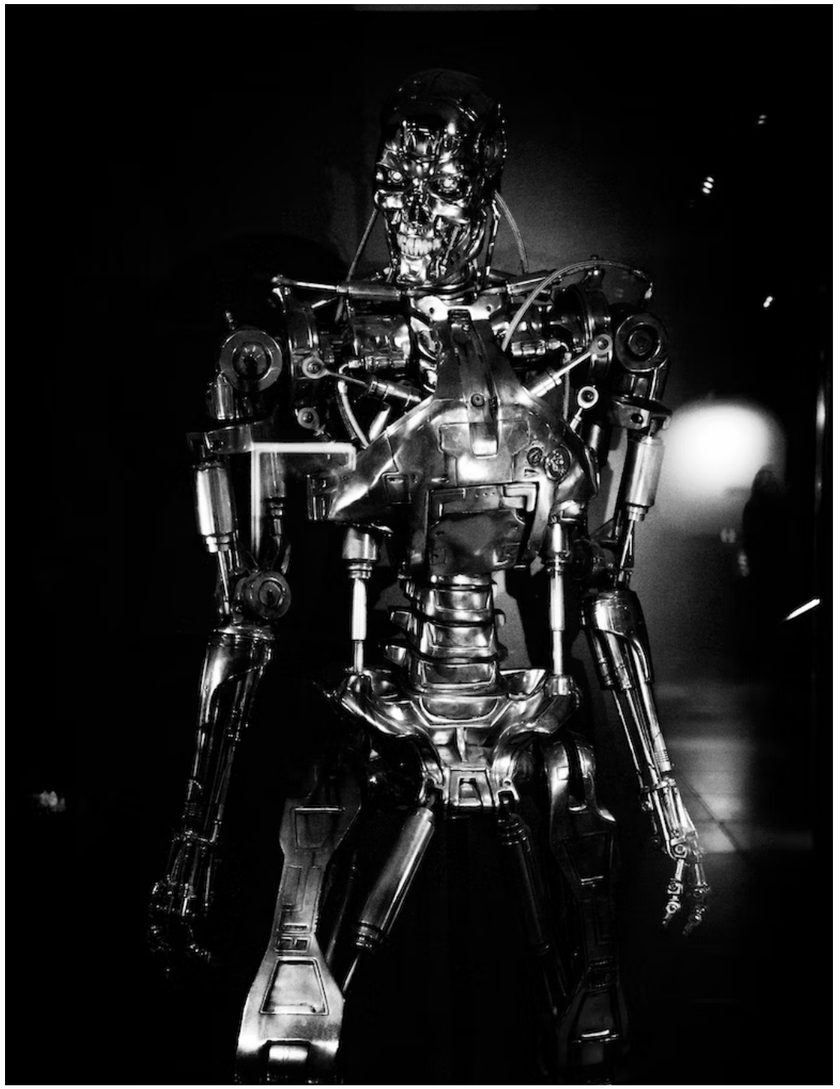
Immense pros

- Productivity
- Research progress
- Engineering solutions
- Companionship and wisdom



Severe cons

- Negative economic disruption
- Malicious use
- Value alignment problems
- Existential catastrophe



The safety debate

AGI can empower

AGI can have negative consequences

Controlling AGI outcomes

Requirements for aligning AGI and human values:

- Clear rules and expectations
- Constructive feedback

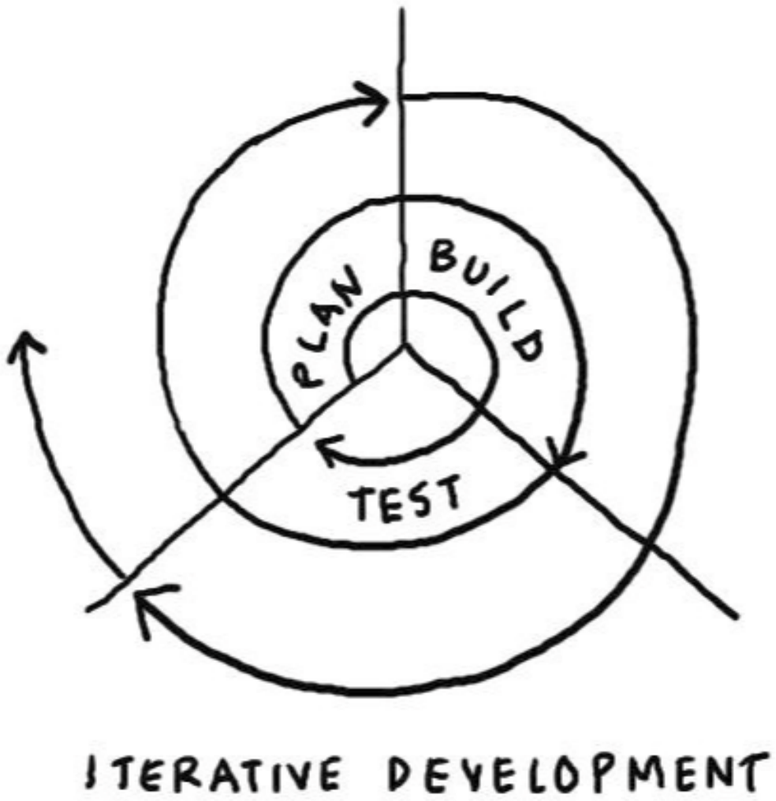
1. Hard constraints
2. Alignment strategies
3. Government intervention

Hard constraints

1. **Boxing** restricts access to the wider world
2. **Interruptibility** adds a stop or off switch

Alignment strategies

- Iterative development
- Constitutional AI
- Multi-stakeholder engagement



[1] Dave Gray

Government intervention

- Beneficial regulations
  - Safety regulations
  - Rules for testing and oversight
  - Transparency standards
- International collaboration

CEOs of various AI companies meeting with UK PM Rishi Sunak in 2023-

[1] UK Prime Minister

Let's practice!

GENERATIVE AI CONCEPTS