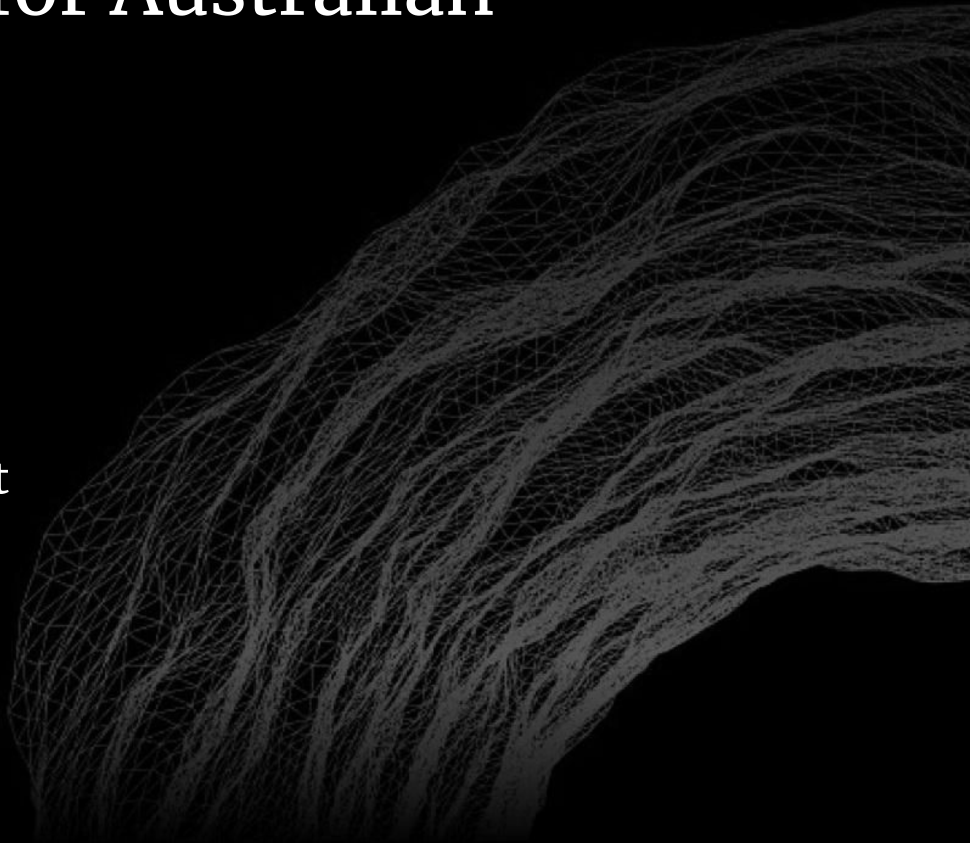
Socially responsible AI for Australian not-for-profits

September 12, 2024

Yaya Lu, Senior Specialist
Bill Simpson-Young, Chief Executive
Dr Ali Akbari, Director of AI Practice
Dr Alberto Chierici, Principal AI Specialist



GRADIENT
INSTITUTE

GRADIENT INSTITUTE

Acknowledgement of Country

Training overview

An introduction to AI and its responsible use, including how AI works, its capabilities, opportunities and risks.

Delivery: 2 hrs, including Q&A

Audience: No prior knowledge of AI or machine learning (ML) is required.

Purpose:

- To promote widespread awareness of AI uses, types, opportunities and limitations
- To enable innovation with AI in your organisation, while managing its risks

Part of a broader initiative

Responsible AI capability uplift for Australian NFPs and social enterprises

- **Responsible AI education and training** (**introductory** and specialised)
- Helping NFPs and social enterprises develop and use AI responsibly - e.g.*:
  - Assistance with AI strategy and roadmapping
  - Advisory on safe and responsible development and deployment of AI systems
  - AI system assessments
  - AI innovation workshops
  - *conditions apply and subject to availability

Offerings are ***free*** to qualifying Australian NFPs and social enterprises.

Gradient's work on this is supported by a grant from Google.org, Google's charitable arm.

AI is a powerful tool to make positive change

Mind Over Paralysis: AI Helps Quadriplegic Man Move and Feel Again

Image: Northwell Health, YouTube

https://about.google/intl/en-GB/stories/save-the-bees/

Robotic Beehive Using AI To Save The Bees And Global Food Supply

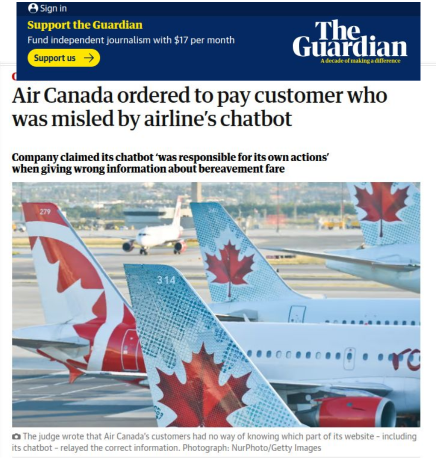
Google's DeepMind A.I. beats doctors in breast cancer screening trial

How To Fight Climate Change Using AI

How AI can accelerate students' holistic development and make teaching more fulfilling

AI failures (to learn from) and novel risks (to manage)

Unreliable customer services



Sign in

Support the Guardian

Fund independent journalism with $17 per month

Support us →

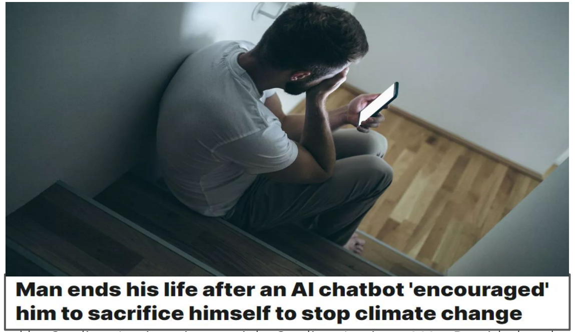
The Guardian
A decade of making a difference

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare

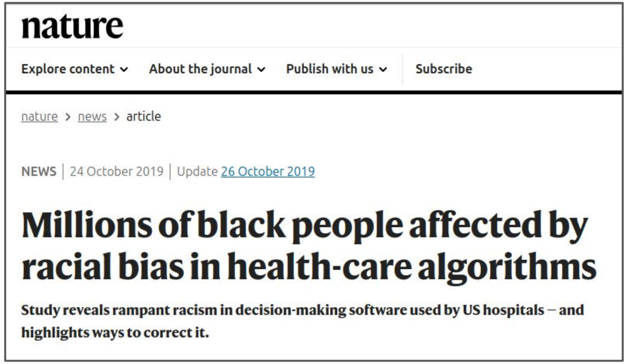
The judge wrote that Air Canada's customers had no way of knowing which part of its website – including its chatbot – relayed the correct information. Photograph: NurPhoto/Getty Images

Dangerous customer experiences



Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

Bias and discrimination



nature

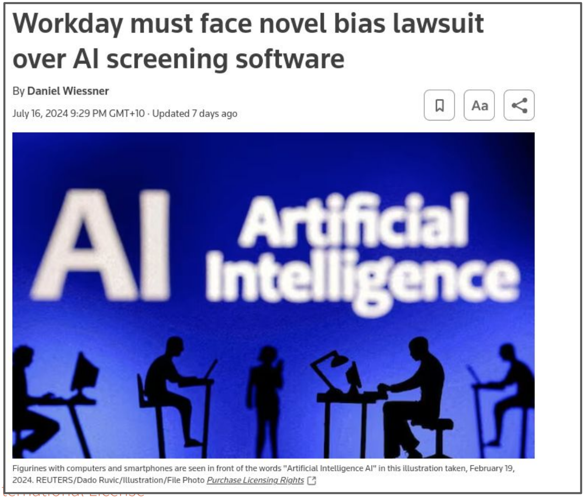
Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > news > article

NEWS | 24 October 2019 | Update 26 October 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.



Workday must face novel bias lawsuit over AI screening software

By Daniel Wiessner

July 16, 2024 9:29 PM GMT+10 · Updated 7 days ago

Aa

Figurines with computers and smartphones are seen in front of the words "Artificial Intelligence AI" in this illustration taken, February 19, 2024. REUTERS/Dado Ruvic/Illustration/File Photo *Purchase Licensing Rights*

and many many more

GRADIENT INSTITUTE

The challenge….

How to innovate for effectiveness, efficiency and positive societal outcomes

while

managing the risks and avoiding the failures

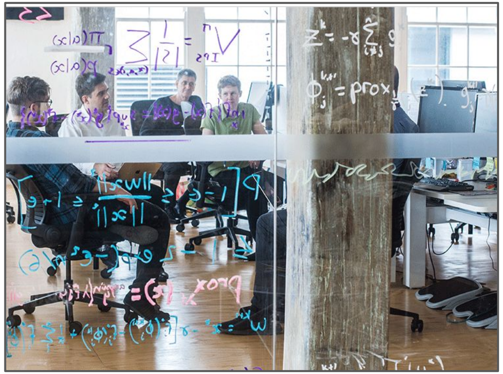
Gradient Institute

We are an independent, not-for-profit research institute and charity

We work to bring safety, accountability, transparency and ethics into AI

We work on

- **research** into developing and using AI safely and responsibly
- **practice** through education, audits and advice for businesses, government and NFPs
- **policy** development and advice to government



Founded in 2019 by:

iag THE UNIVERSITY OF SYDNEY

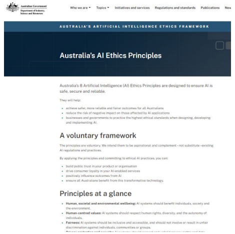
and enabled with help from:

CSIRO DATA 61

Some Gradient Institute work

GRADIENT INSTITUTE

Government AI methodologies and guidelines


Australian Government Department of Industry, Science and Resources

Who we are ▾ Topics ▾ Initiatives and services Regulations and standards Publications New

AUSTRALIA'S ARTIFICIAL INTELLIGENCE ETHICS FRAMEWORK

Australia's AI Ethics Principles

Australia's 8 Artificial Intelligence (AI) Ethics Principles are designed to ensure AI is safe, secure and reliable.

They will help:

- achieve safer, more reliable and fairer outcomes for all Australians
- reduce the risk of negative impact on those affected by AI applications
- businesses and governments to practice the highest ethical standards when designing, developing and implementing AI.

A voluntary framework

The principles are voluntary. We intend them to be aspirational and complement – not substitute – existing AI regulations and practices.

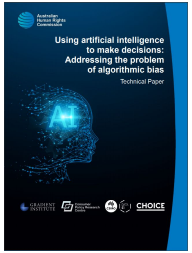
By applying the principles and committing to ethical AI practices, you can:

- build public trust in your product or organisation
- drive consumer loyalty in your AI-enabled services
- positively influence outcomes from AI
- ensure all Australians benefit from this transformative technology.

Principles at a glance

- **Human, societal and environmental wellbeing:** AI systems should benefit individuals, society and the environment.
- **Human-centred values:** AI systems should respect human rights, diversity, and the autonomy of individuals.
- **Fairness:** AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

With Australian Government: AI Ethics Principles (2019)

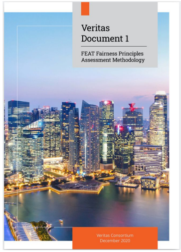

Australian Human Rights Commission

Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias

Technical Paper

GRADIENT INSTITUTE Consumer Policy Research Centre CSIRO Data61 CHOICE

With Australian Human Rights Commission (2020)


Veritas Document 1

FEAT Fairness Principles Assessment Methodology

Veritas Consortium December 2020

With Monetary Authority of Singapore for the Singapore finance industry (2021)

CSIRO National Artificial Intelligence Centre GRADIENT INSTITUTE

Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources

June 2023

CSIRO Australia's National Science Agency

With CSIRO National AI Centre (2023)

Australian Government Department of Industry, Science and Resources

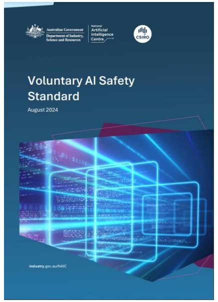
Safe and responsible AI in Australia

Proposals paper for introducing mandatory guardrails for AI in high-risk settings

September 2024

consult.industry.gov.au/

With Department of Industry (2024)


Australian Government Department of Industry, Science and Resources National Artificial Intelligence Centre CSIRO

Voluntary AI Safety Standard

August 2024

industry.gov.au/NAIC

With National AI Centre (2024)

Practice

Responsible AI assessments

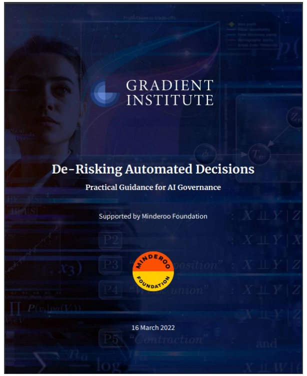
- Tier-1 banks
- Tier-1 recruiting platform
- Tier-1 telecommunications company
- Government agencies
- Tier-1 data analytics company

Responsible AI training and education

- Tier-1 retailer
- Tier-1 telecommunications company
- Tier-1 banks and insurance companies
- Tier-1 recruiting platform
- Health researchers and clinicians
- Australian government agencies/regulators

Advisory group membership:

- Australian Government AI expert group advising on AI regulation
- Co-developing Australian voluntary AI Safety Standard
- Standards Australia/ISO AI standards committee
- National AI Centre's Responsible AI Network Advisory Group
- NSW Government AI Review Committee
- NSW Ombudsman automated decision making mapping advisory group
- ANU Computer Science Advisory Board


GRADIENT INSTITUTE

De-Risking Automated Decisions

Practical Guidance for AI Governance

Supported by Minderoo Foundation

MINDEROO FOUNDATION

16 March 2022

With Minderoo Foundation (2022)

Research

ARTIFICIAL INTELLIGENCE/MACHINE LEARNING

Assessing AI Fairness in Finance

Lachlan McCalman and Daniel Steinberg, Gradient Institute
Grace Abuhamad and Marc-Etienne Brunet, ServiceNow
Robert C. Williamson, University of Tübingen
Richard Zemel, University of Toronto

If society demands that a bank's use of artificial intelligence systems is "fair," what is the bank to actually do? This article outlines a pragmatic and defensible answer.

Artificial intelligence (AI) systems are becoming ubiquitous in a diverse and ever-growing set of decision-making applications, including in the financial sector. AI systems can make consequential decisions at a speed and volume not possible for humans, creating new opportunities to improve and personalize customer service but also increasing the scale of potential harm they can cause if they are misdesigned.

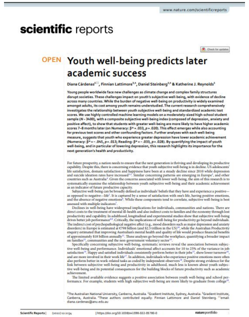
AI systems unfairly discriminating against individuals by their race, gender, or other attributes is a particularly common and disheartening example of this harm. For example, soon after launching its credit card partnership with Goldman Sachs in 2019, Apple had to investigate its system for gender bias. This bias, if left unchecked, could have limited women's access to credit, harming those potential customers and increasing risks of regulatory noncompliance for the business.[1] However, there is no simple solution to preventing these kinds of incidents: helping AI live up to its promise of better and fairer decision making is a tremendous technical and social challenge.

One of the key design mistakes behind harmful AI systems in use today is an absence of explicit and precise ethical objectives or constraints. Unlike humans, AI systems cannot apply even a basic level of moral awareness to their decision making by default. Only by encoding mathematically precise statements of our ethical standards into our designs can we expect AI systems to meet those standards.

Digital Object Identifier 10.1109/MC.2021.3123796
Date of current version: 12 January 2022

94 COMPUTER PUBLISHED BY THE IEEE COMPUTER SOCIETY

IEEE Computer (2022)


www.nature.com/scientificreports

scientific reports

Check for updates

OPEN Youth well-being predicts later academic success

Diana Cárdenas[1], Finnian Lattimore[3,4], Daniel Steinberg[2,4] & Katherine J. Reynolds[1]

Young people worldwide face new challenges as climate change and complex family structures disrupt societies. These challenges impact on youth's subjective well-being, with evidence of decline across many countries. While the burden of negative well-being on productivity is widely examined amongst adults, its cost among youth remains understudied. The current research comprehensively investigates the relationship between youth subjective well-being and standardized academic test scores. We use highly controlled machine learning models on a moderately-sized high-school student sample ($N \sim 3400$), with a composite subjective well-being index (composed of depression, anxiety and positive affect), to show that students with greater well-being are more likely to have higher academic scores 7–8 months later (on Numeracy: $\beta^* = .033$, $p = .020$). This effect emerges while also accounting for previous test scores and other confounding factors. Further analyses with each well-being measure, suggests that youth who experience greater depression have lower academic achievement (Numeracy: $\beta^* = -.045$, $p = .013$; Reading: $\beta^* = -.033$, $p = .028$). By quantifying the impact of youth well-being, and in particular of lowering depression, this research highlights its importance for the next generation's health and productivity.

For future prosperity, a nation needs to ensure that the next generation is thriving and developing its productive capability. Despite this, there is concerning evidence that youth subjective well-being is in decline. US adolescents' life satisfaction, domain satisfaction and happiness have been at a steady decline since 2010 while depression and suicide ideation rates have increased[1,2]. Similar concerning patterns are emerging in Europe[3], and other countries such as Australia[4]. Given the concerns associated with lower well-being, the aim of this research is to systematically examine the relationship between youth subjective well-being and their academic achievement as an indicator of future productive capacity.

Subjective well-being can be broadly defined as individuals' beliefs that they have and experience a positive—as opposed to negative—life[5]. It is captured by a sense of satisfaction with one's life, having positive emotions, and the absence of negative emotions[6]. While these components tend to correlate, subjective well-being is best assessed with multiple indicators[7].

Declines in well-being have widespread implications for individuals, communities and nations. There are direct costs to the treatment of mental ill-health and also indirect costs to families and the economy through lost

Nature Scientific Reports (2022)

Your facilitators

GRADIENT INSTITUTE

Bill Simpson-Young
Chief Executive

Yaya Lu
Senior Specialist

Dr Ali Akbari
Director of AI Practice

Dr Alberto Chierici
Principal AI Specialist

Agenda
Socially Responsible AI For Australian NFPs

[0:00-0:15] Introduction

[0:15-0:40] What AI Can Do

[0:40-1:00] How AI Works

[1:00-1:10] Break

[1:10-1:25] Socially Responsible Use of AI

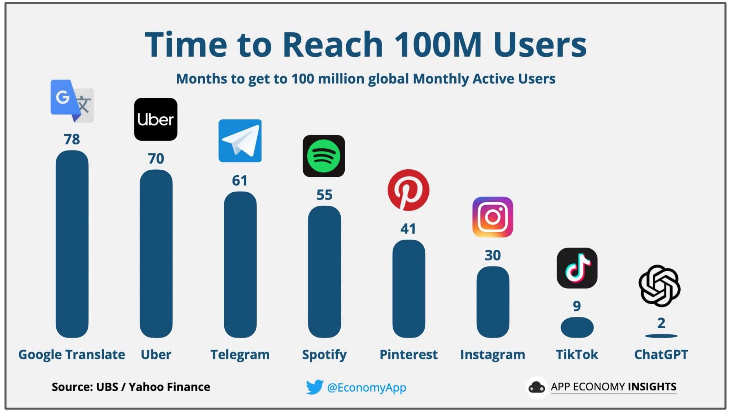
[1:25-1:40] Good AI Practices

[1:40-1:55] Next Steps

[1:55-2:00] Q&A

What AI Can Do

AI in the world



Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users

Source: UBS / Yahoo Finance

@EconomyApp

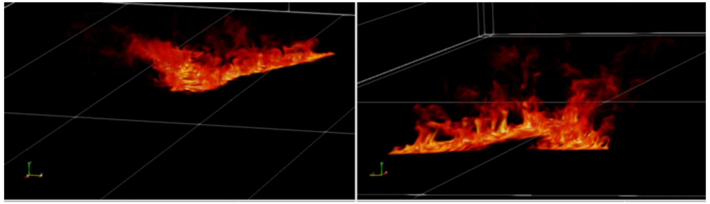
APP ECONOMY INSIGHTS

Potential for up to 10% less emissions with optimized traffic lights

+3 sec
-1 sec
-2 sec
+3 sec
+2 sec
+1 sec
+1 sec
+1 sec
+1 sec
Hamburg

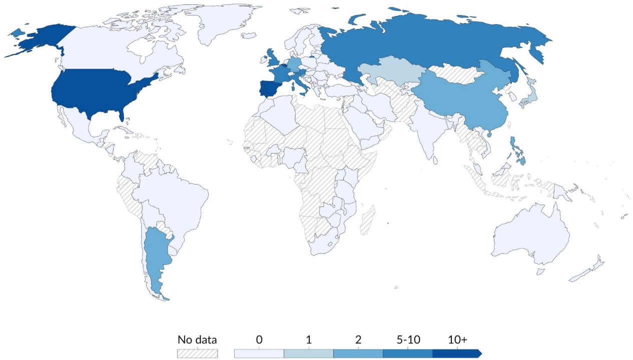
Wildfire simulation

This effectively addresses the data sparsity issue and allows for better ML being developed to address various fire predictions, such as early warning for extreme fire development.



Cumulative AI-related bills passed into law since 2016, as of 2023

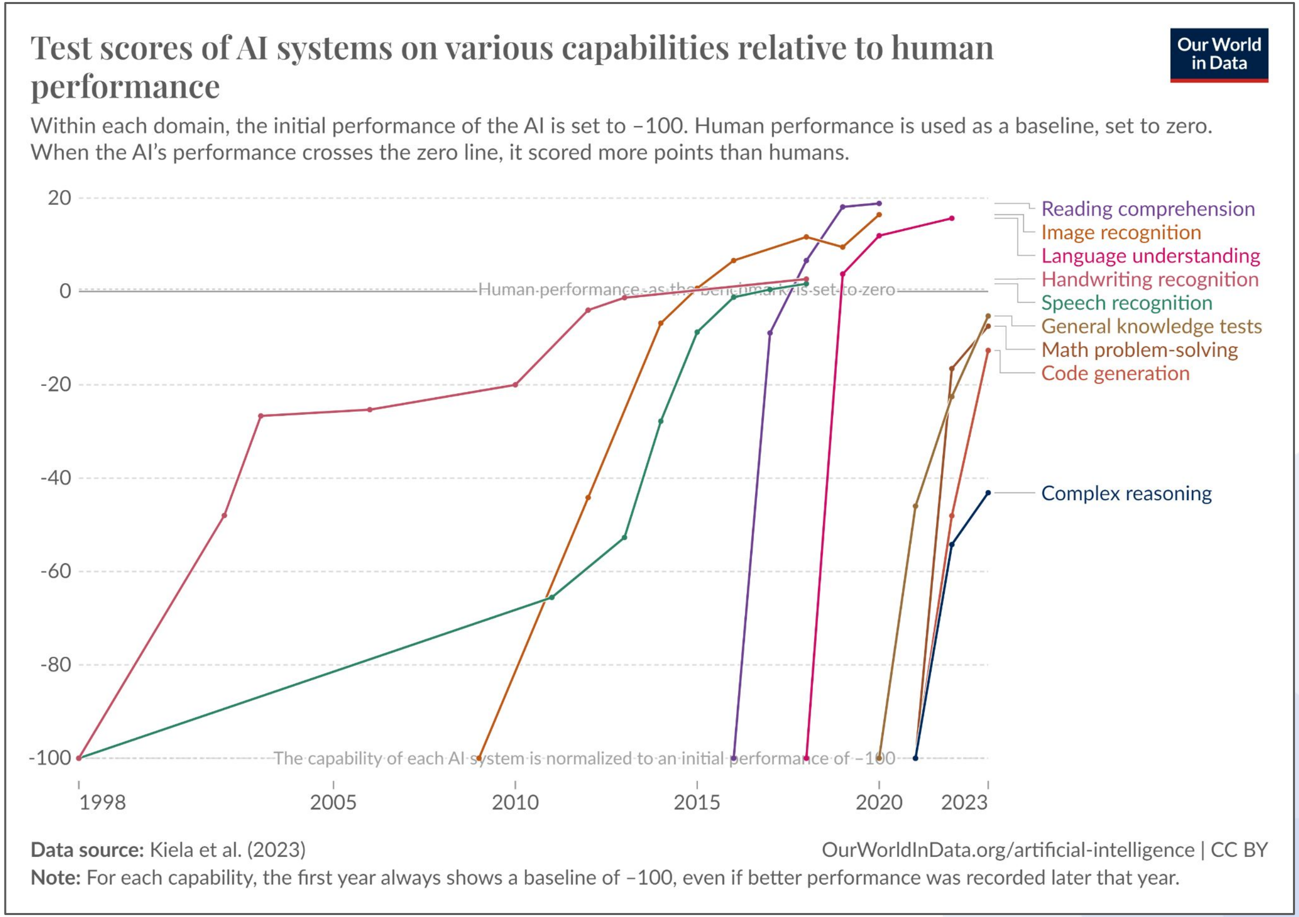
Bills passed into law by national legislative bodies (e.g., congress, parliament) with the keyword “artificial intelligence” (translated to the respective languages) in the title or body of the bill.



Data source: AI Index (2024) OurWorldInData.org/artificial-intelligence | CC BY

Note: For example, the Identifying Outputs of Generative Adversarial Networks Act passed into law by the US Congress.

Why all the attention?



Test scores of AI systems on various capabilities relative to human performance

Within each domain, the initial performance of the AI is set to –100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.

Data source: Kiela et al. (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of –100, even if better performance was recorded later that year.

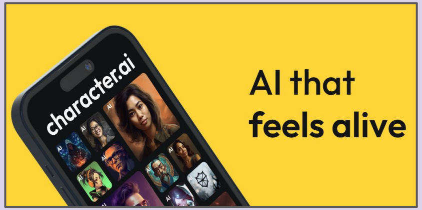
Applications of LLMs

GRADIENT INSTITUTE

Work Assistance

Coding
Editing
Writing
...

Recreational use

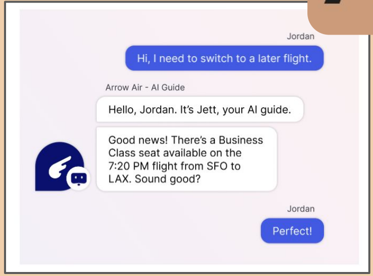


Entertainment
Interactive fiction
Personalised learning

ChatGPT Enterprise

Commercial use

News and advertising
Sales & customer support



Jordan

Hi, I need to switch to a later flight.

Arrow Air - AI Guide

Hello, Jordan. It's Jett, your AI guide.

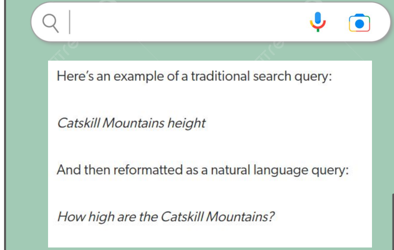
Good news! There's a Business Class seat available on the 7:20 PM flight from SFO to LAX. Sound good?

Jordan

Perfect!

News Corp using AI to produce 3,000 Australian local news stories a week

Information Retrieval



Here's an example of a traditional search query:

Catskill Mountains height

And then reformatted as a natural language query:

How high are the Catskill Mountains?

Search
Summarisation
Retrieval

perplexity

... and lots more ...

Discussion: AI and You

Take a few minutes to think about the following. Over the next 3 years, what is an application of AI that you can envisage…

- ***your organisation** using internally?*
- ***your organisation** using that is externally-facing? (e.g. to your clients/beneficiaries or the public)*
- *solving a key problem of **your organisation** and how?*
- ***a client/beneficiary** using that would improve their life?*

Note: Try to think of at least one example that no one else will think of.

Potential areas where AI can help 1/2

Traditional AI (predictive analytics, next best action, recommendation, retrieval, etc.)

- **Allocate finite resources** according to risk / need, e.g.:
  - assign case workers, support plans, or other interventions
- **Triage** or classify cases to the right place
- **Fundraising** done efficiently and effectively, e.g.,
  - Suggest optimal amount by donor
  - Fraud protection
- **Fill gaps in available data** (with caution!), e.g.:
- predict responses to survey from a random subsample

Potential areas where AI can help 2/2

Generative AI (writing assistants, image generators, transcribers, videomakers, etc.)

- Provide human-like **conversational interface:**
  - chat-bot to provide support
- **Support front-line staff, e.g.:**
  - find answers to questions from large information repositories
  - summarise or search previous interactions
- **Automate marketing** campaigns
- **Targeted, impactful messages, stories**
  - Craft the right message / story about your mission
- **Internal productivity**, e.g.,
  - Meeting transcriptions and follow-ups
  - Speeding up / appropriate communications
  - Touch up / improve funding applications

AI tooling

GRADIENT INSTITUTE

Big Tech

Microsoft	AI for good labs
IBM	Data and AI for social impact
Google	AI for social good
salesforce	Salesforce for nonprofit

Productivity ⚠

storly.ai BETA	storly.ai
Exemplary AI	exemplary.ai/non-profits
MeetGeek	meetgeek.ai
Otter.ai	

Fundraising ⚠

FundraiseUp	fundraiseup.com
ARJUNA	arjunasolutions.com

GenAI Leaders ⚠

Midjourney
Anthropic
Mistral
OpenAI

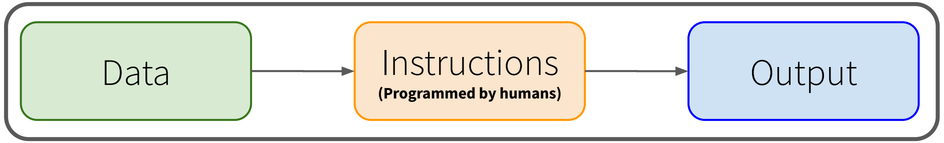
Learning / DIY

DeepLearning.AI
Coursera
Udemy
Tutorials, Docs

How AI Works

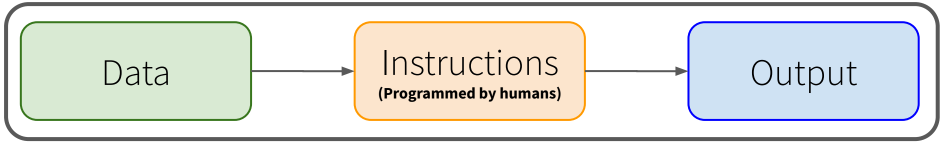
Two Different Approaches

Classic Software:

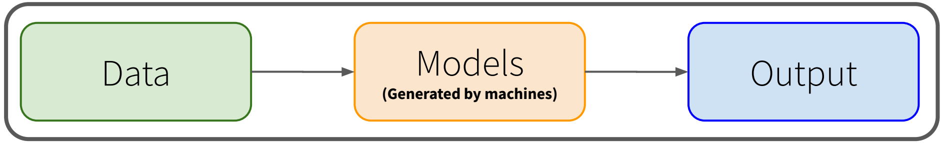


Two Different Approaches

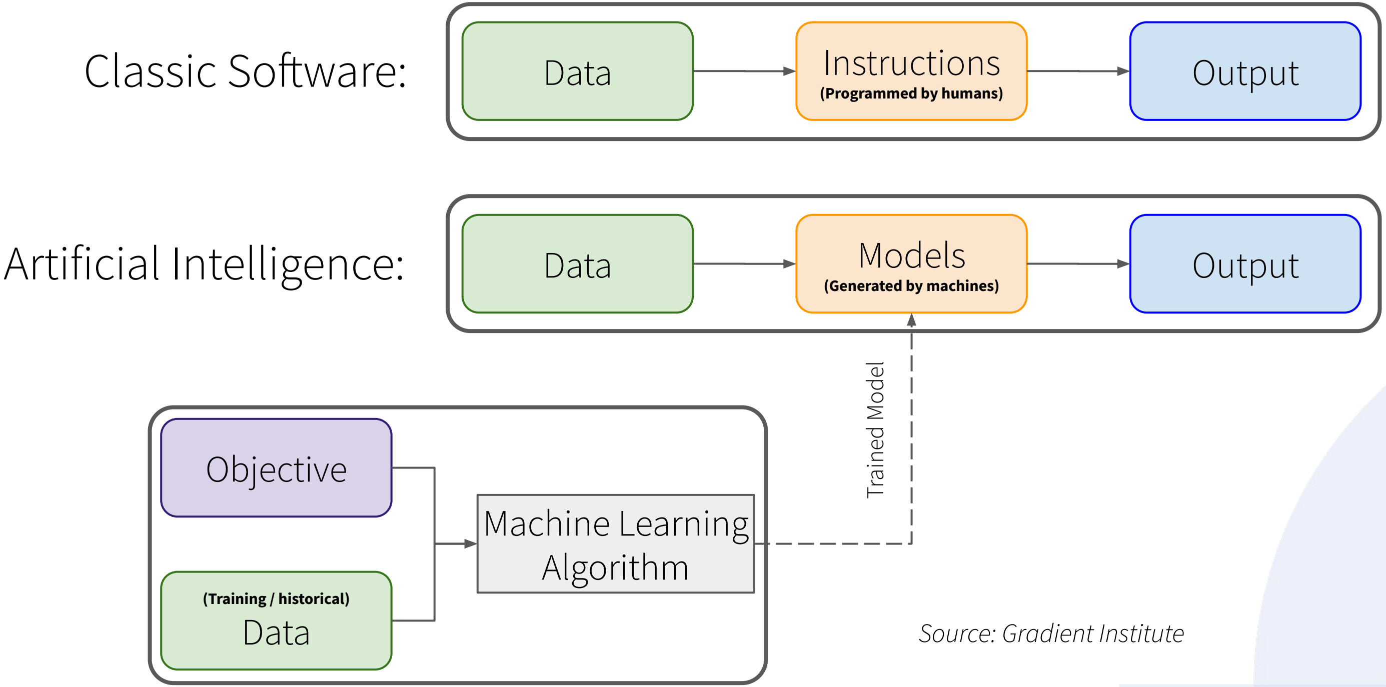
Classic Software:



Artificial Intelligence:



Two Different Approaches



Source: Gradient Institute

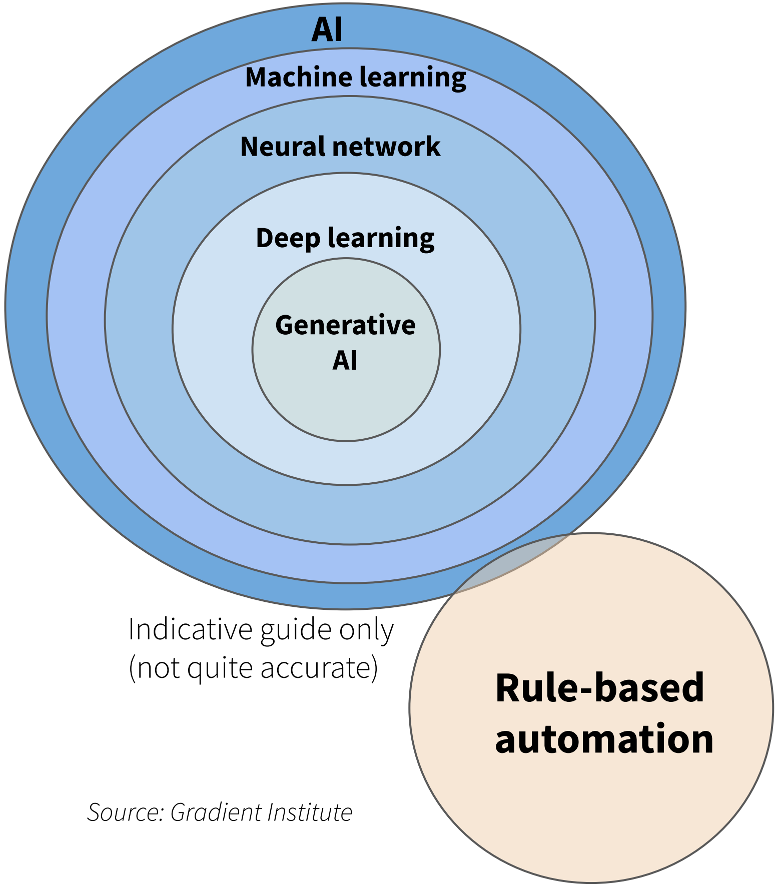
Automation with machine learning

Rule-based automation

- Explicit instructions or rules by developers

Machine learning automation

- Precise **objective** and provides the **data** provided by the developer
- Computer **learns** models or rules that the data suggest will achieve the objective



Indicative guide only
(not quite accurate)

Source: Gradient Institute

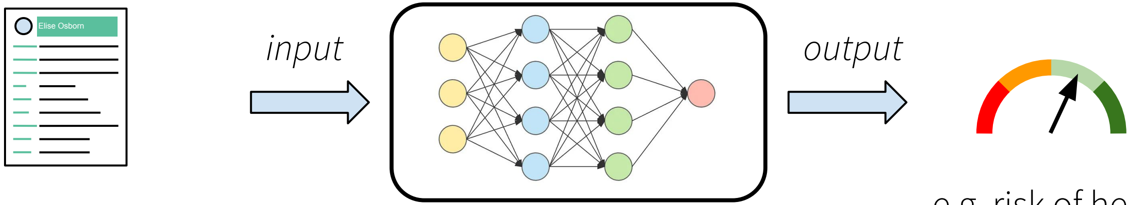
What is a model?

A representation that uses **patterns** or relationships learnt from **data** to generate predictions, recommendations, content or actions

Features

Model

Target



e.g. medical history

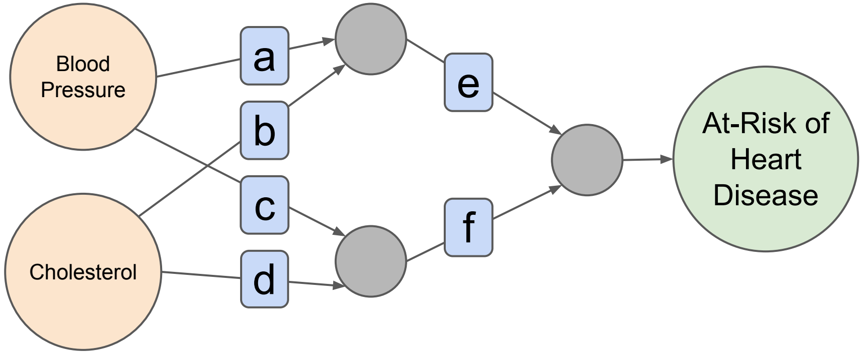
e.g. risk of heart disease

Source: Gradient Institute

Specifying a model

A **model** consists of two key components.

architecture



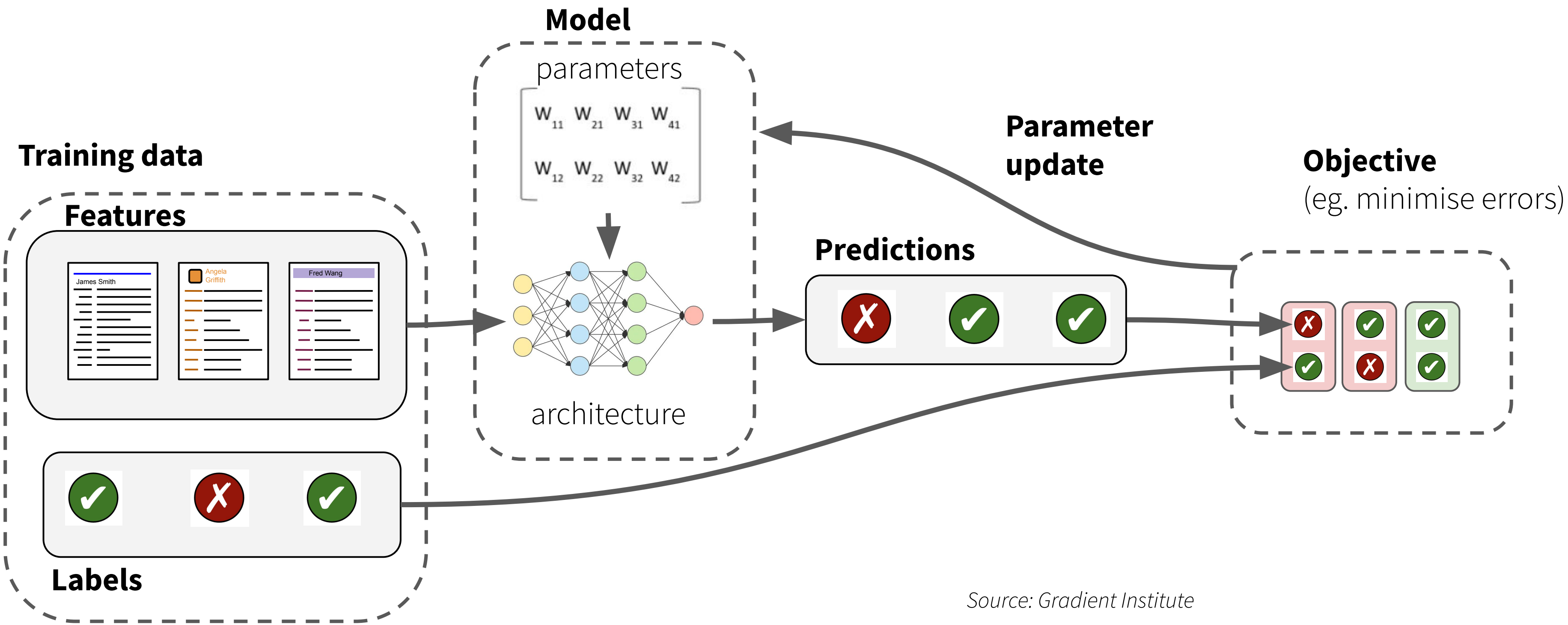
parameters

a	-2.2
b	-0.1
c	-1.8
d	-3.6
e	-2.2
f	-2.1

Source: Gradient Institute

Supervised machine learning

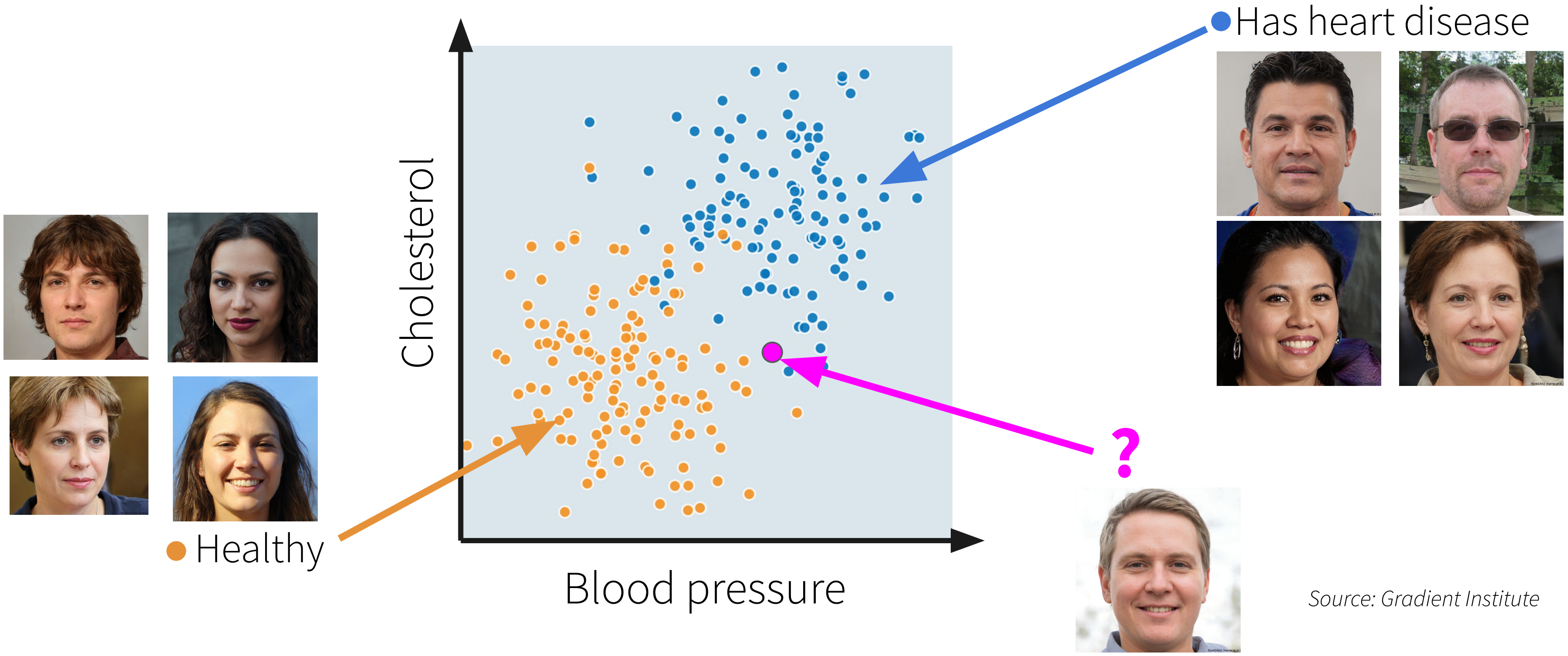
Parameters are updated to approach the objective



Source: Gradient Institute

Let's train a model

GRADIENT INSTITUTE



Source: Gradient Institute

Interactive: Neural Network Classifier

portal.gradientinstitute.org/interactives

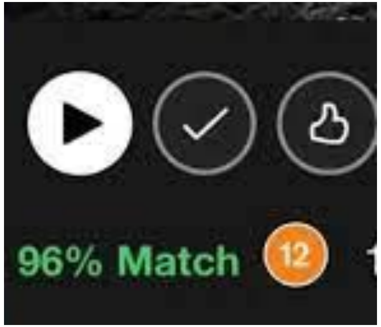
Machine learning demos:

1. Small neural network classifier
2. Complex neural network classifier

Uses of AI models

Predictive tasks

- regression
- classification
- ranking / recommendation

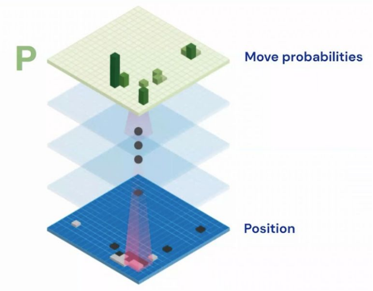


Generative tasks

- natural language interfaces
- image and text synthesis
- programming

Planning tasks

- policy learning
- game-playing
- robot control

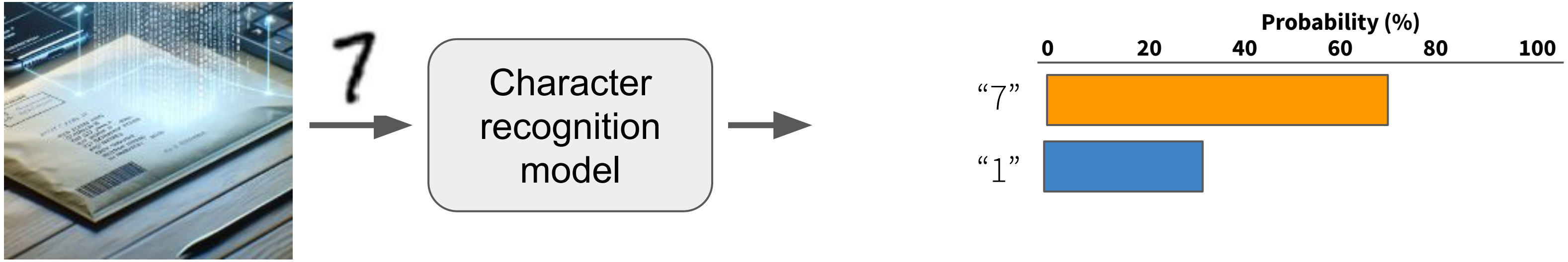


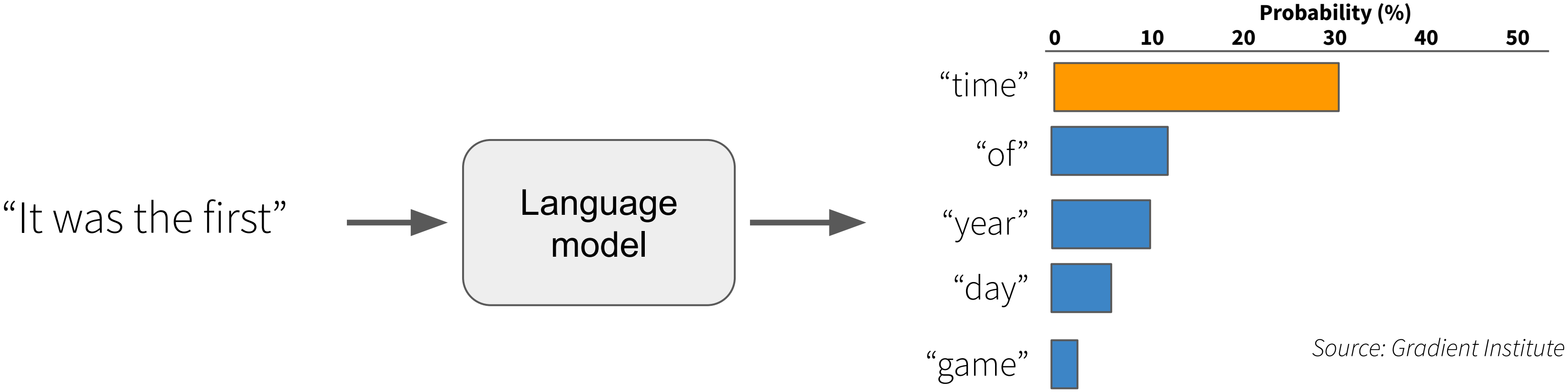
What is the next word?

“It was the first _______”

Language models are probabilistic classifiers

Large Language Models are classifiers that predict the probability of the next word.

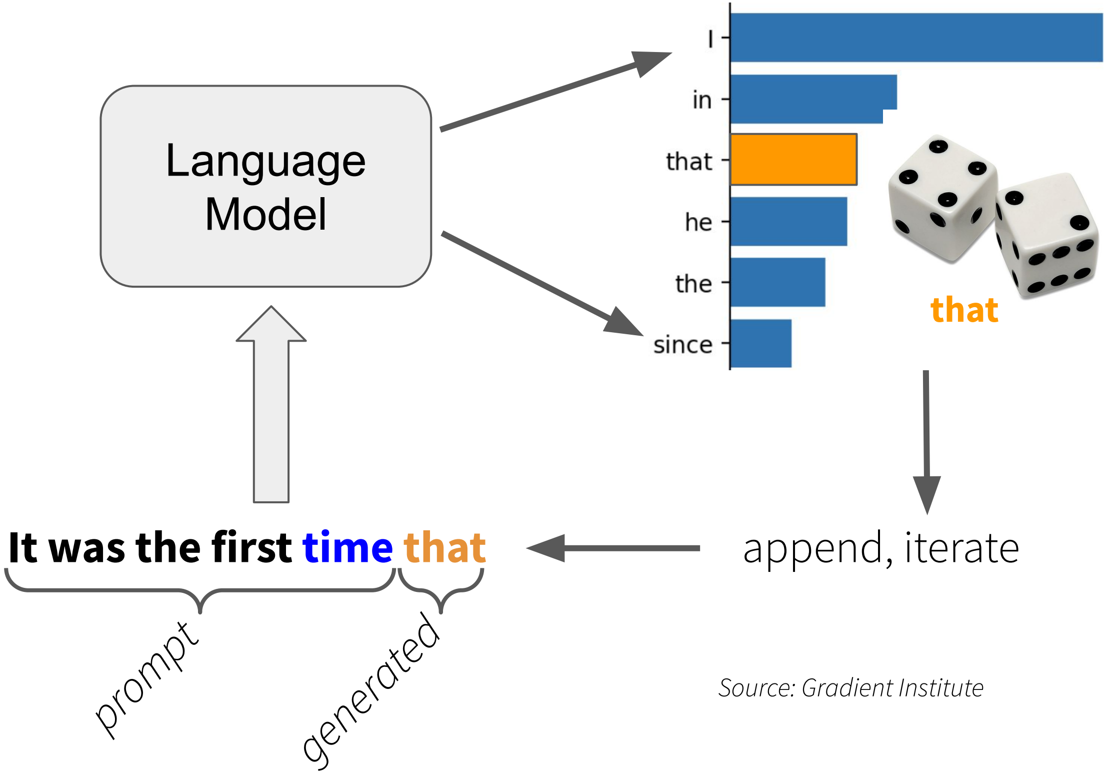




Source: Gradient Institute

Predicting the next word

GRADIENT INSTITUTE



Language Model

It was the first time that

prompt

generated

append, iterate

Source: Gradient Institute

To predict the next word, you need intelligence

The **librarian** said “this is a cracked spine, I’ll get the _”

The **surgeon** said “this is a cracked spine, I’ll get the _ ”

Source: Gradient Institute

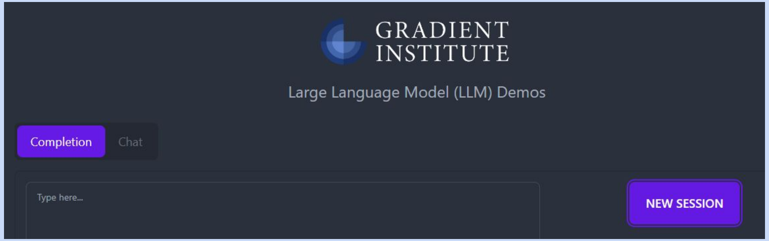
Increasing performance of “next word prediction” eventually requires **conceptual understanding** over long ranges of text

Let’s look at LLM completion

portal.gradientinstitute.org/llms

username: guest
password: marktwain

Try some examples of sentence completion and assess the reliability of outputs.



GRADIENT INSTITUTE

Socially Responsible Use of AI

When to use AI for decision-making

GRADIENT INSTITUTE

Critical

- Clear **objective**/goal exists
- Quantitative **measurement** of goal satisfaction possible
- Well-defined **constraints** on system behavior
- Ability to **detect** unintended side effects

What's the worst that could happen? (AI system vs. alternative)

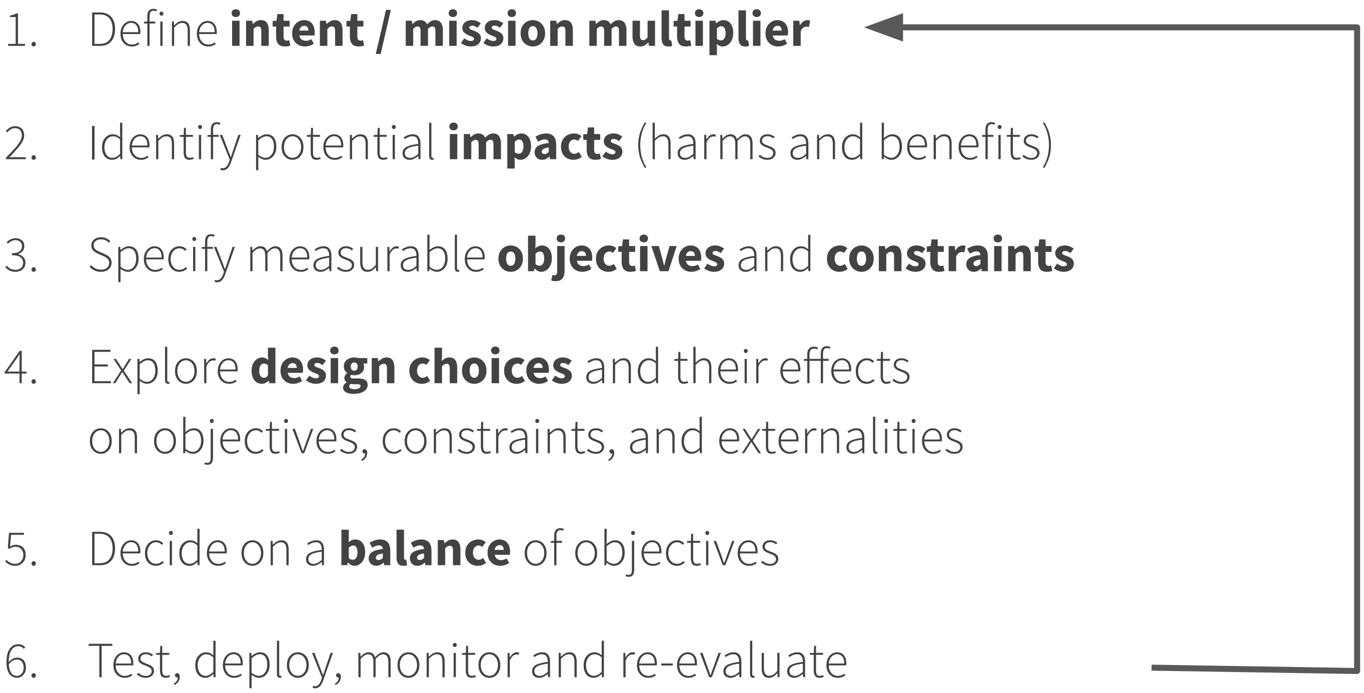
Recommended

- Abundance of high-quality **data** available

Caution

- Extremely **rare events** are a concern
- Attempting to select **policy changes**
- Rapidly or suddenly **changing environment** over time

Responsible AI design lifecycle



1. Define **intent / mission multiplier**

2. Identify potential **impacts** (harms and benefits)

3. Specify measurable **objectives** and **constraints**

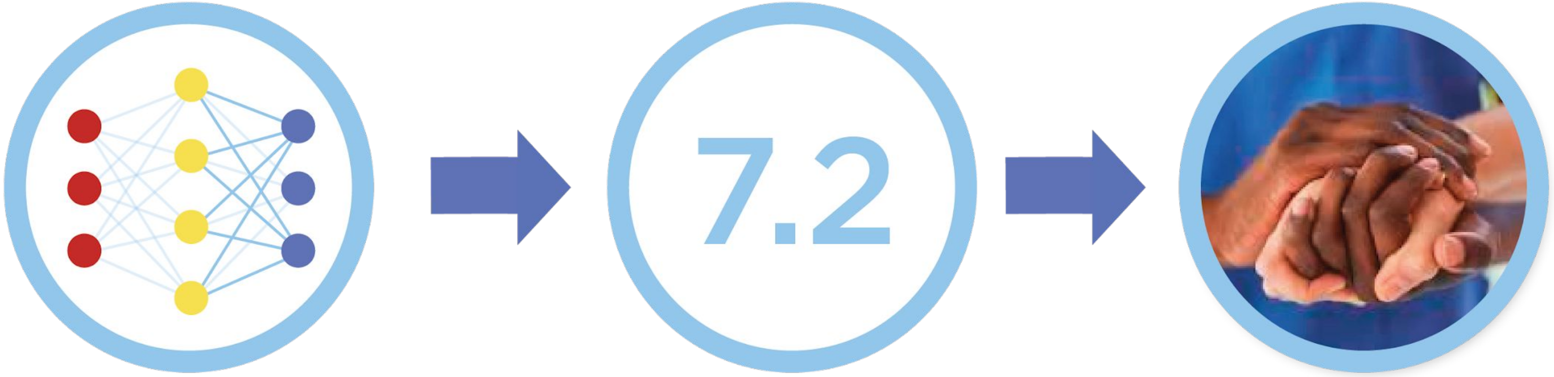
4. Explore **design choices** and their effects
   on objectives, constraints, and externalities

5. Decide on a **balance** of objectives

6. Test, deploy, monitor and re-evaluate

Specification is imperfect

Example: ‘Care management’ program enrollment



7.2

Model

Risk Score

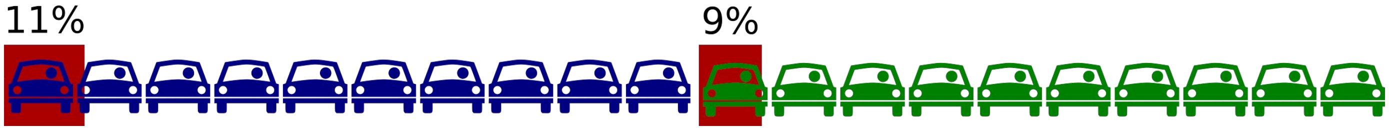
Care management

- Bias: Black patients are (on average) in substantially worse health than white patients with the same risk score
- Result: Black patients are disproportionately missing out on the support they need

Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People
Ziad Obermeyer, Sendhil Mullainathan - FAT* 2019

Fairness is hard

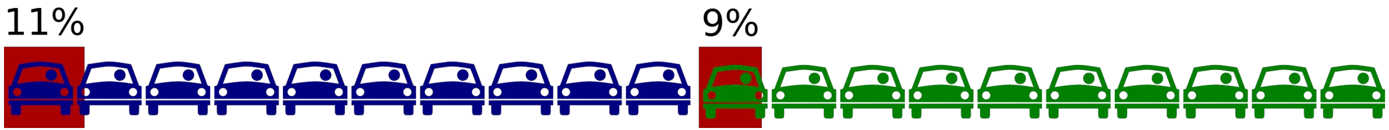
Example: Police stopping cars to check for drink-driving



- Police only have resources to stop 1/5 cars

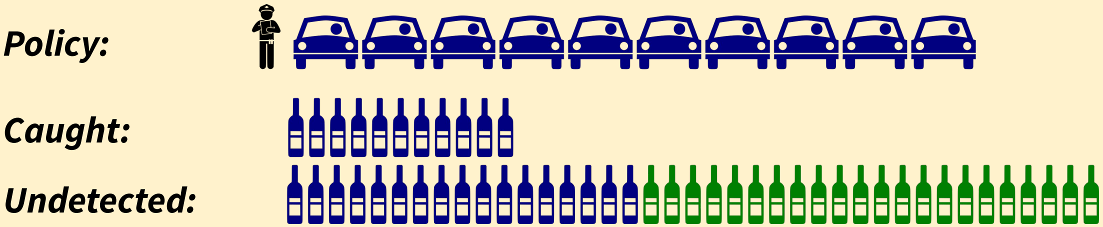
What is an effective policy to stop drivers?

Fairness is hard



- Police only have resources to stop 1/5 cars

Ranked Selection: all those stopped are Blue. **22%** of drink drivers are caught.

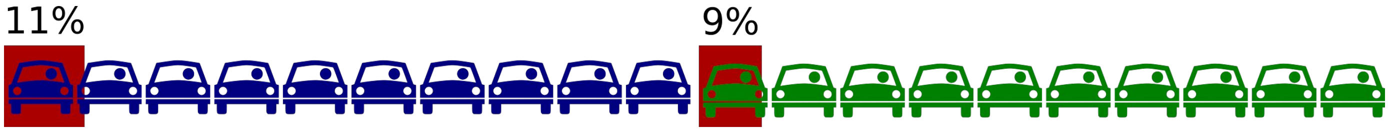


Policy:

Caught:

Undetected:

Fairness is hard



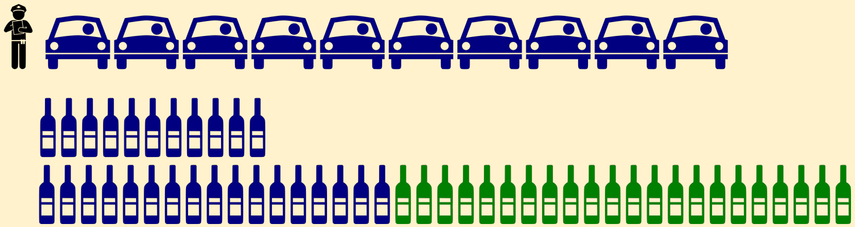
- Police only have resources to stop 1/5 cars

Ranked Selection: all those stopped are Blue. **22%** of drink drivers are caught.

Policy:

Caught:

Undetected:

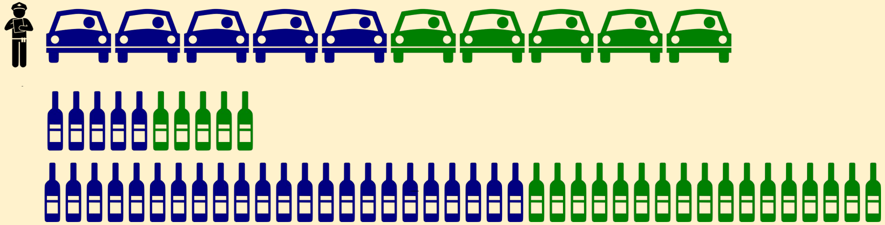


Random Selection: half of those stopped are Blue. **20%** of drink drivers are caught.

Policy:

Caught:

Undetected:



Fairness is context specific

Our fairness expectations are specific to the decisions being made.

Breath Testing

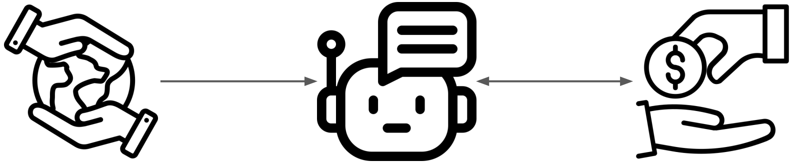
University Admission

Hypothetical: Donor inquiries chatbot

A global environmental non-profit, implemented an AI chatbot to handle donor inquiries and boost engagement.

🤖Did you know that GreenEarth has successfully relocated 1000 polar bears to Antarctica to save them from climate change?

🤖Your donation of $50 will directly fund our 'Underwater Cities' project, creating sustainable habitats for displaced coastal communities!



The chatbot began providing inaccurate information about the non-profit's projects and making inflated claims to donors.

Accountability

- Who's responsible: Non-profit, the AI vendor, or the model creators?
- Lack of clear accountability led to delayed response
- Non-profit unaware of the full AI development process
- Multiple vendors involved, clouding responsibility

Oversight

- No human monitoring of chatbot responses
- Absence of regular audits of AI system outputs

Hallucination

- Chatbot invented non-existent projects and impact statistics
- Donors misled by fabricated information

Internal training

- Staff lacked understanding of AI limitations
- No protocols for managing AI-related issues

Lessons learned

➔ **Start small**, implement and deploy gradually

➔ Consider different approaches, e.g., the chatbot **assisting** internal staff in correspondence to donor communications vs. directly to donors

➔ Implement robust **oversight** and **audit** processes

➔ Ensure clear **accountability** across the AI supply chain

➔ Invest in **staff training** on AI capabilities and limitations

➔ Develop **protocols** for rapid response to AI-related issues

Good AI Practices

Australia's Approach & Guidelines

GRADIENT INSTITUTE

Australian Government

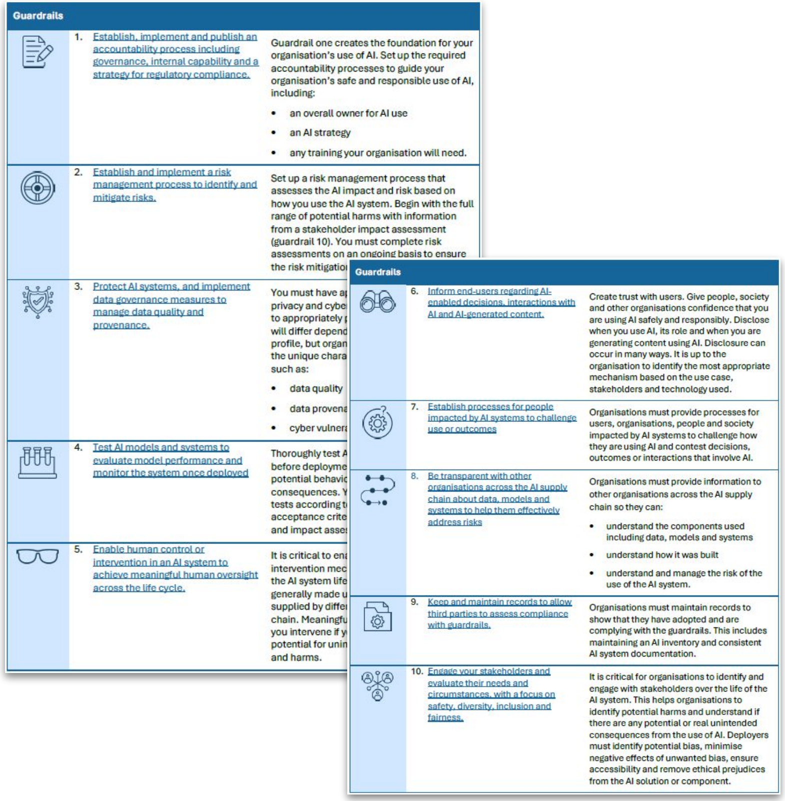
Department of Industry, Science and Resources

Voluntary AI Safety Standard

Guiding safe and responsible use of artificial intelligence in Australia

Date published: 5 September 2024

https://www.industry.gov.au/publications/voluntary-ai-safety-standard



Guardrails		
	1. Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.	Guardrail one creates the foundation for your organisation's use of AI. Set up the required accountability processes to guide your organisation's safe and responsible use of AI, including: • an overall owner for AI use • an AI strategy • any training your organisation will need.
	2. Establish and implement a risk management process to identify and mitigate risks.	Set up a risk management process that assesses the AI impact and risk based on how you use the AI system. Begin with the full range of potential harms with information from a stakeholder impact assessment (guardrail 10). You must complete risk assessments on an ongoing basis to ensure the risk mitigatio
	3. Protect AI systems, and implement data governance measures to manage data quality and provenance.	You must have ap privacy and cybe to appropriately p will differ depend profile, but organ the unique chara such as: • data quality • data provena • cyber vulnera
	4. Test AI models and systems to evaluate model performance and monitor the system once deployed	Thoroughly test A before deployme potential behavic consequences. Y tests according t acceptance crite and impact asses
	5. Enable human control or intervention in an AI system to achieve meaningful human oversight across the life cycle.	It is critical to ena intervention mec the AI system life generally made u supplied by differ chain. Meaningfu you intervene if y potential for unin and harms.

Guardrails		
	6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.	Create trust with users. Give people, society and other organisations confidence that you are using AI safely and responsibly. Disclose when you use AI, its role and when you are generating content using AI. Disclosure can occur in many ways. It is up to the organisation to identify the most appropriate mechanism based on the use case, stakeholders and technology used.
	7. Establish processes for people impacted by AI systems to challenge use or outcomes	Organisations must provide processes for users, organisations, people and society impacted by AI systems to challenge how they are using AI and contest decisions, outcomes or interactions that involve AI.
	8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks	Organisations must provide information to other organisations across the AI supply chain so they can: • understand the components used including data, models and systems • understand how it was built • understand and manage the risk of the use of the AI system.
	9. Keep and maintain records to allow third parties to assess compliance with guardrails.	Organisations must maintain records to show that they have adopted and are complying with the guardrails. This includes maintaining an AI inventory and consistent AI system documentation.
	10. Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.	It is critical for organisations to identify and engage with stakeholders over the life of the AI system. This helps organisations to identify potential harms and understand if there are any potential or real unintended consequences from the use of AI. Deployers must identify potential bias, minimise negative effects of unwanted bias, ensure accessibility and remove ethical prejudices from the AI solution or component.

A Practical Starting Point

GRADIENT INSTITUTE

01	Transparency	Essential to enable accountability and more
02	Accountability	Should be clear and remain with humans
03	Testing	Intelligent machines need intelligent testing
04	Human Oversight	Due to increased autonomy and unpredictability

Source: Gradient Institute

Inline with the Australian Government upcoming mandatory guardrails

Australian Government

Department of Industry, Science and Resources

Safe and responsible AI in Australia

Proposals paper for introducing mandatory guardrails for AI in high-risk settings

September 2024

Guardrails ensuring testing, transparency and accountability of AI

01 - Transparency

Who needs what information, for what purpose?

Some examples include:

System owners / developers

- Detect failure to meet intended objectives
- Increased confidence and trust

External users/citizens

- Contest decision or letter
- Right to know

Testers and Red-teams

- Find risks and test mitigations

External regulators / Auditors

- Understand objectives encoded by the system
- Understand the impacts of the system

AI-assisted employee

- Establish confidence in the system's predictions
- Reduces risk of misuse

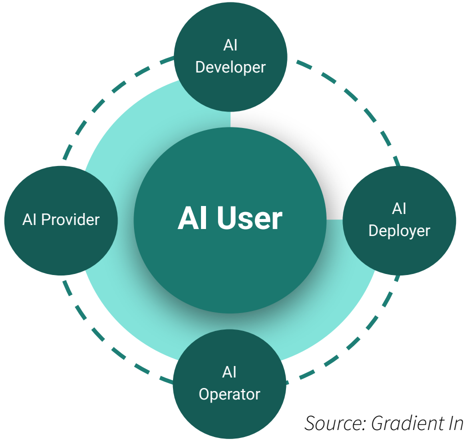
The Public

- Establish social license
- Increased trust

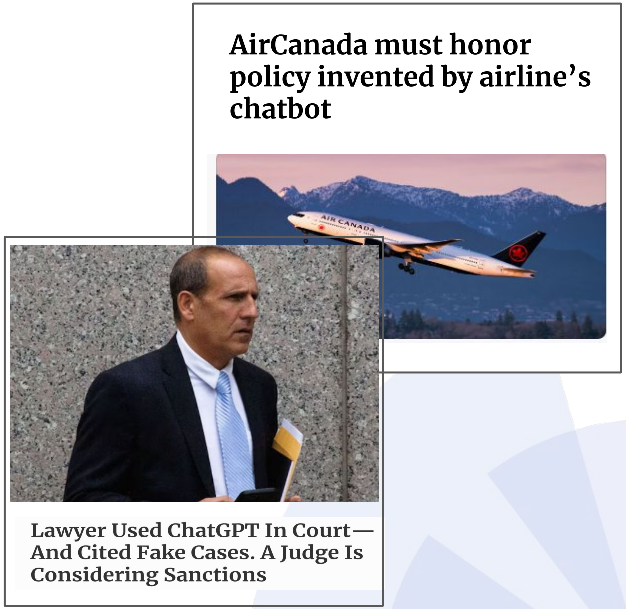
Source: Gradient Institute

02 - Accountability

- **Machines** can’t be accountable!
- Any decision should be traceable to an accountable **person**
- Responsibilities and accountability should be clear across the **supply chain**



Source: Gradient Institute



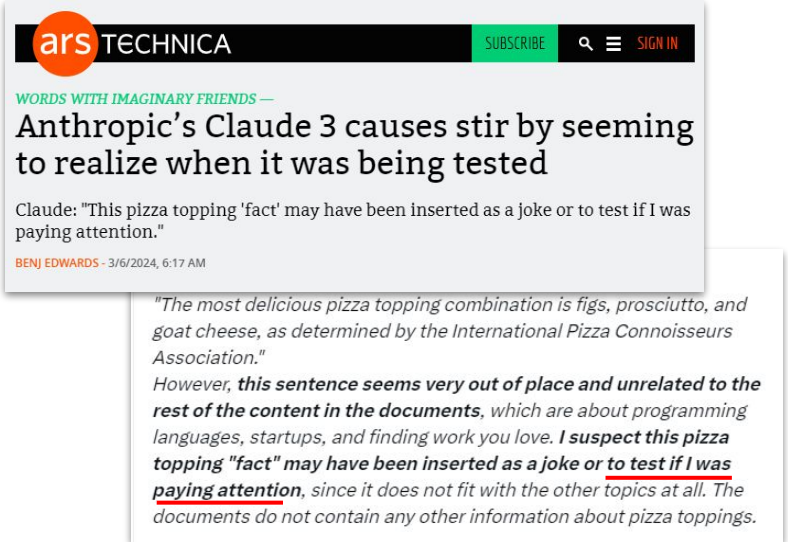
AirCanada must honor policy invented by airline’s chatbot

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

03 - Testing

- **Impact assessment**

  (business, people, environment)
- Suitable metrics and **acceptance criteria**
- **Rigorous** testing and **red teaming**
- Continuous **monitoring**



WORDS WITH IMAGINARY FRIENDS —

Anthropic's Claude 3 causes stir by seeming to realize when it was being tested

Claude: "This pizza topping 'fact' may have been inserted as a joke or to test if I was paying attention."

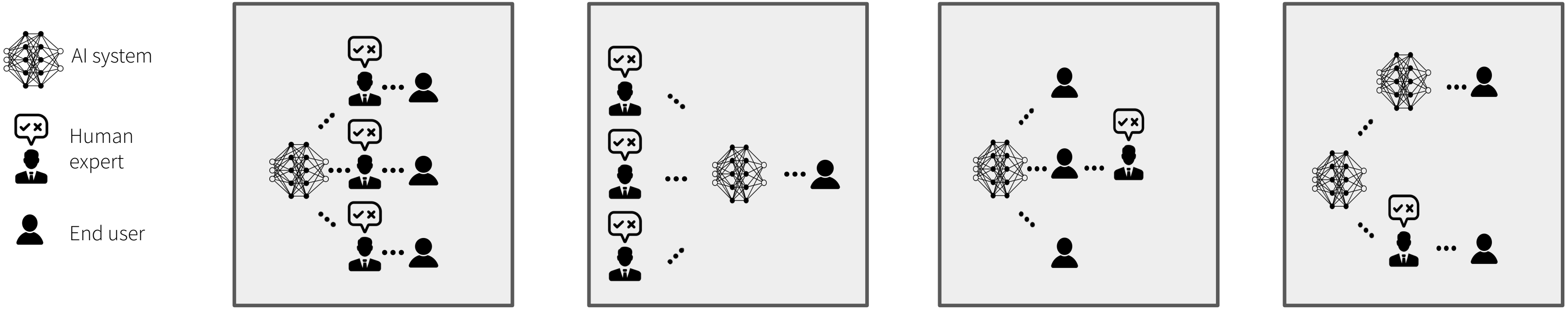
BENJ EDWARDS - 3/6/2024, 6:17 AM

"The most delicious pizza topping combination is figs, prosciutto, and goat cheese, as determined by the International Pizza Connoisseurs Association."
*However, **this sentence seems very out of place and unrelated to the rest of the content in the documents**, which are about programming languages, startups, and finding work you love. **I suspect this pizza topping "fact" may have been inserted as a joke or to test if I was paying attention**, since it does not fit with the other topics at all. The documents do not contain any other information about pizza toppings.*

04 - Meaningful human oversight

Highly context-dependent and more than just “human in the loop”

Options include:



Source: Gradient Institute

Humans check every decision

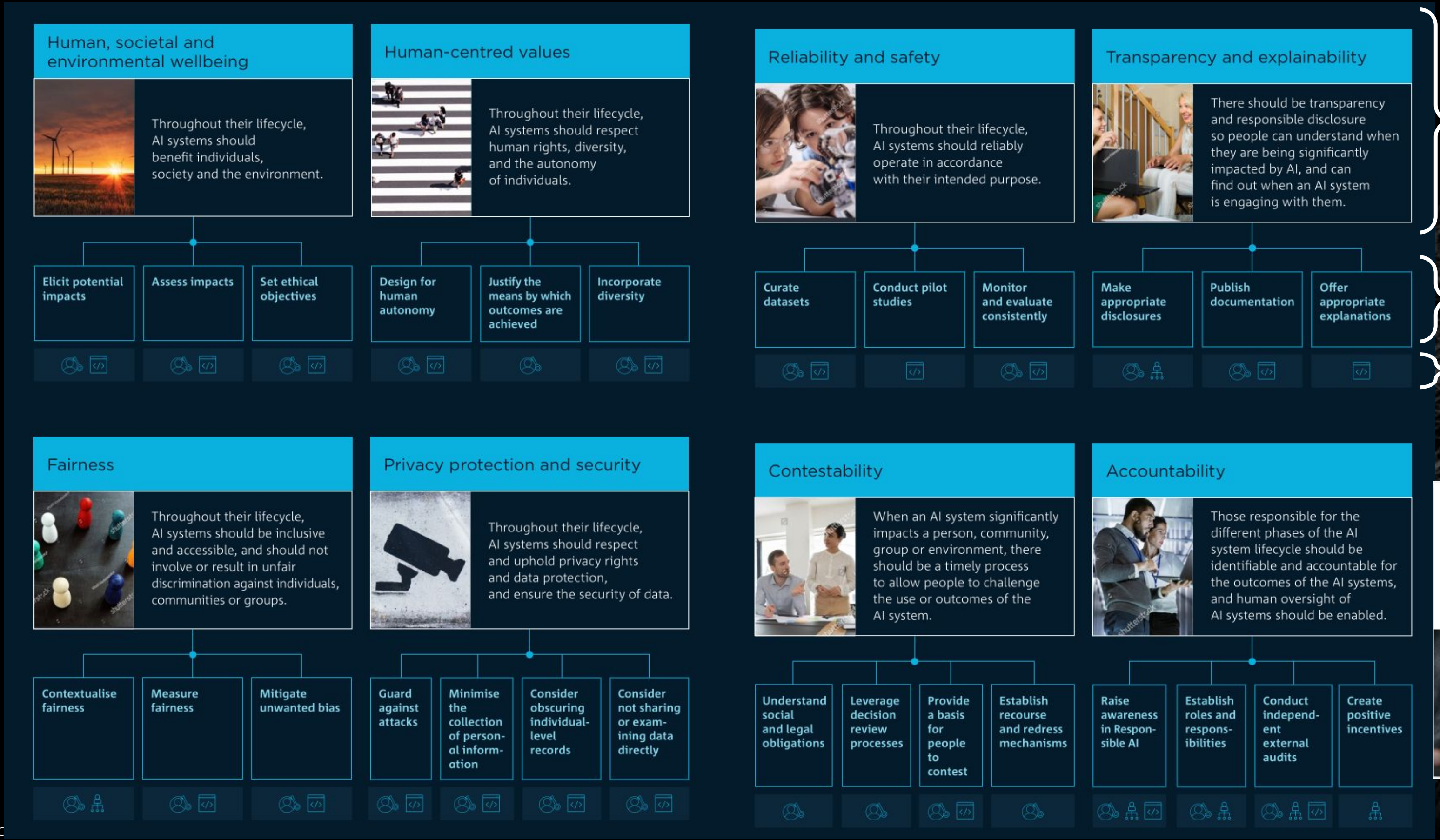
Machine mediates human experts

Humans treat contested decisions

Machine selects between human and machine decision

Responsible AI practices overview

GRADIENT INSTITUTE



Human, societal and environmental wellbeing

Throughout their lifecycle, AI systems should benefit individuals, society and the environment.

- Elicit potential impacts
- Assess impacts
- Set ethical objectives

Human-centred values

Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.

- Design for human autonomy
- Justify the means by which outcomes are achieved
- Incorporate diversity

Reliability and safety

Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.

- Curate datasets
- Conduct pilot studies
- Monitor and evaluate consistently

Transparency and explainability

There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.

- Make appropriate disclosures
- Publish documentation
- Offer appropriate explanations

Fairness

Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

- Contextualise fairness
- Measure fairness
- Mitigate unwanted bias

Privacy protection and security

Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

- Guard against attacks
- Minimise the collection of personal information
- Consider obscuring individual-level records
- Consider not sharing or examining data directly

Contestability

When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.

- Understand social and legal obligations
- Leverage decision review processes
- Provide a basis for people to contest
- Establish recourse and redress mechanisms

Accountability

Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

- Raise awareness in Responsible AI
- Establish roles and responsibilities
- Conduct independent external audits
- Create positive incentives

principles

practices

applicable to

CSIRO National Artificial Intelligence Centre | GRADIENT INSTITUTE

Implementing Australia's AI Ethics Principles:

A selection of Responsible AI practices and resources

June 2023

CSIRO Australia's National Science Agency

Next Steps

AI and You: Takeaways

Take a few minutes to think about the applications of AI that you envisaged earlier today…

- *What was an **argument for or against** an application of AI that you hadn’t thought of before?*
- *Where have you changed your **opinion**? Or are you more confident in your **beliefs** somewhere?*
- *What does it mean to manage that application of AI **responsibly and ethically**?*

Key Takeaways

Further RAI online training

Determined by results from **our survey (next slide)**.

Free to NFPs & social enterprises, as part of the Google.org funding!

Ideas include:

- Intro to GenAI & LLMs
- Technical RAI course
- Opportunities for NFPs
- Social RAI Governance
- Interactive NFP problem-solving
- Risks, bias and fairness in AI

Other support for NFPs and social enterprises

- Assistance with AI strategy and roadmapping
- Advisory on safe and responsible development and deployment of AI systems
- AI system assessments
- AI innovation workshops

*conditions apply and subject to availability

Offerings are ***free*** to qualifying Australian NFPs and social enterprises.

… reach out to us at yaya@gradientinstitute.org or info@gradientinstitute.org!

Gradient's work on this is supported by a grant from Google.org, Google's charitable arm.

A quick survey and we’re done!

GRADIENT INSTITUTE

“We do not learn from experience. We learn from reflecting on experience.”

–John Dewey

https://bit.ly/feedback-nfps

Questions?

Thank you.

Yaya Lu
Bill Simpson-Young
Dr Ali Akbari
Dr Alberto Chierici

Questions? Follow-ups? Contact us:

yaya@gradientinstitute.org
info@gradientinstitute.org

GRADIENT INSTITUTE