ProductDock

Ethics of AI

Friday talk

About me.

- UX/UI Designer since 2014
- Studied New Technologies of Communication
- Based in Oporto
- Joined PD Portugal in 2021

UX & AI.

Working with AI tools

How to integrate AI tools in UX process?

Which manual work can be automated?

How will AI impact design jobs in the future?

Designing AI solutions

How AI solutions change user interactions?

How to make AI solutions user friendly?

How to design good AI?

AI ETHICS

Agenda.

01
What is AI Ethics

02
Why does it matter

03
Core principles

04
Final notes

What is AI ethics?

What is AI ethics.

Ethics is a set of **moral principles** aimed to help us discern between right and wrong.

Ai ethics is a multidisciplinary field that studies how to **optimize Ai's beneficial impact** while reducing risks and adverse consequences.

AI ethics questions/dilemas.

- How to get rid of **AI discrimination**?
- How to **prevent harm**?
- How can we make **AI decisions understandable**?
- How to **protect AI from hackers**?
- How much **control** should AI have?
- Who's **responsible** for AI's mistakes?
- How can AI respect and **promote human rights**?
- ...

Why does it matter?

Benefits vs concerns.

FORBES > INNOVATION > AI

The Synergy Of Humans And AI Is Reshaping The Workforce For The Futu

Neil Sahota Contributor

Neil Sahota is a globally sought after speaker and business advisor.

The future work has humans and AI working side-by-side. GETTY

Artificial intelligence (AI) has emerged as a transformative force,

Malicious use of AI could cause 'unimaginable' damage, says UN boss

António Guterres calls for new UN body along the lines of IPCC to tackle threats posed by artificial intelligence

António Guterres said the advent of generative AI could be a defining moment for disinformation and hate speech. Photograph: Ralph Tedy Erol/Reuters

Malicious use of artificial intelligence systems could cause a "horrific" amount of death and destruction, the UN secretary general has said, calling for a new UN body to tackle the threats posed by the technology.

'Time is running out': can a future of undetectable deepfakes be avoided?

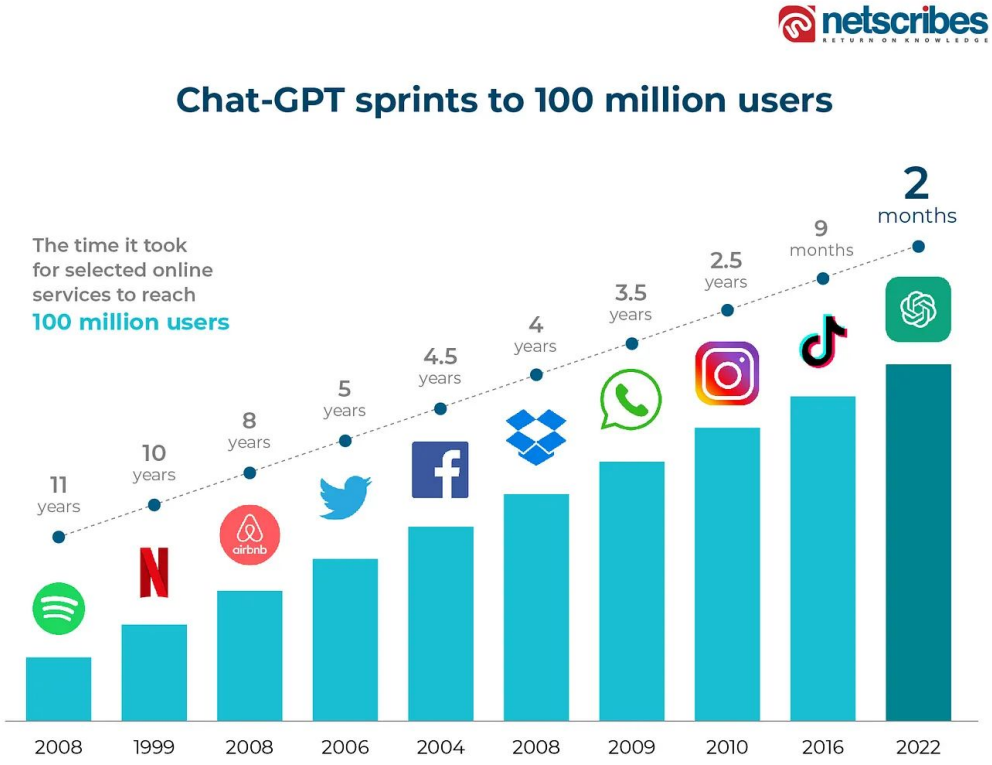
signs of generative AI images are disappearing as ology improves, and experts are scrambling for hods to counter disinformation

AI content is outpacing the human eye and finding and removing it automatically is raph: Maria Korneeva/Getty Images

ith more than 4,000 shares, 20,000 comments, and 100,000 reactions on Facebook, the photo of the elderly woman, sitting behind her homemade 122nd birthday cake, has unquestionably gone viral. "I started decorating cakes from

Social media example.

It’s all going too fast.



Regulation.

MENU

Topics › Digital › Artificial intelligence ›

EU AI Act: first regulation on artificial intelligence

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023

Last updated: 18-06-2024 - 16:29

6 min read

Table of contents

- AI Act: different rules for different risk levels
- Transparency requirements
- Supporting innovation
- Next steps
- More on the EU's digital measures

Link

What are the big companies doing?

Google's AI principles

AWS responsible AI

IBM's principles for trust and transparency

Meta responsible AI

Core principles.

Core principles.

01

Fairness and non-discrimination

How to get rid of AI bias?

02

Non-maleficence

How to prevent harm?

03

Transparency and explainability

Understanding what AI does

04

Accountability and responsibility

Who's responsible for AI's mistakes?

05

Privacy and data protection

How to manage personal data?

06

Safety and security

How to protect users?

Fairness and non-discrimination.

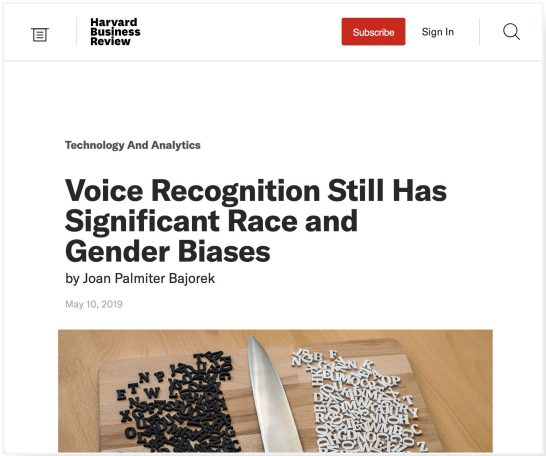
What the principle says.

Fairness and non-discrimination

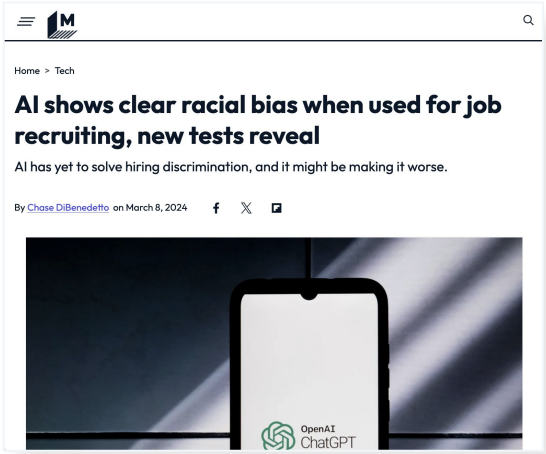
AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

Consequences.

Link



Link



Link

Consequences.

Bias

Bias is an inclination, prejudice, preference or tendency towards or against a person, group, thing, idea or belief.

Biases are usually unfair and are often based on stereotypes, rather than knowledge or experience.

Bias can be intentional or unintentional and can lead to unfairness, inequality, and distorted understanding of situations or groups of people.

3 categories of AI bias.

Systemic bias

Procedures and practices of particular institutions that result in certain groups being advantaged or disadvantaged

Statistical Computational Bias

Occur when sample is not representative of the population

Human bias

Systematic errors in thinking based on simple mental shortcuts

How to mitigate AI bias.

- Focus on values
- Use representative data
- Use clean data
- Transparency
- With AI

Use AI to fight AI.

Farilearn

An open source toolkit from Microsoft for data scientists and developers to assess and improve the fairness of their AI systems.

Accenture's AI testing services

Rely on a "Teach and Test" methodology to train AI systems to avoid biases.

Bias Analyzer

A cloud-based application by PwC that flags potential biases in AI outputs.

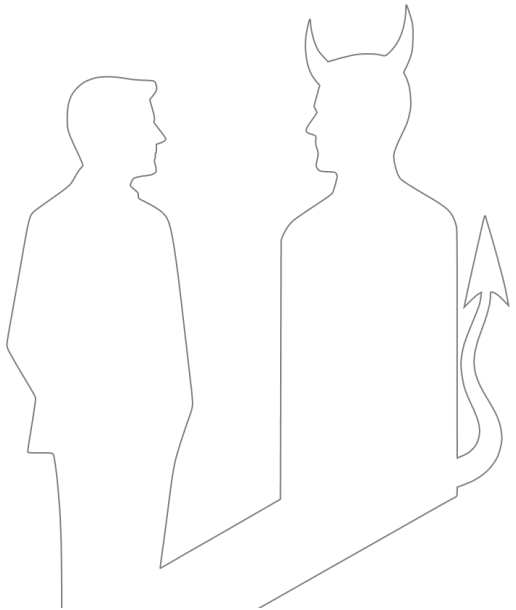
FarilML

A toolbox developed by MIT student Julius Adebayo for auditing predictive models by analyzing the model's inputs.

Google's What-If tool

This tool questions what fairness means and allows product developers to sort the data according to a different type of fairness, allowing humans to see the trade-offs in different ways to measure fairness and make decisions accordingly.

Non-maleficence.



What the principle says.

Non-maleficence

Do no harm. The use of AI systems must not go beyond what is necessary to achieve a legitimate aim. Risk assessment should be used to prevent harms which may result from such uses.

Consequences.

From Deepfakes to Malware: AI's Expanding Role in Cyber Attacks

Mar 19, 2024 Newsroom

Large language models (LLMs) powering artificial intelligence (AI) tools today could be exploited to develop self-augmenting malware capable of bypassing YARA rules.

Hackers from China, Russia, Iran Use AI Tools, Microsoft Says

Illustration/Diálogo

BY JULIETA PELCASTRE/DIÁLOGO
MARCH 18, 2024

According to Microsoft and OpenAI, hackers are using artificial intelligence (AI) systems like ChatGPT to improve their cyberattacks. In joint research published on

Singapore

AI fuelling more sophisticated phishing attempts, cyberattacks

Malicious actors are likely to benefit as AI continues to improve and be adopted, said the Cyber Security Agency.

AI has also allowed malicious actors to scale up their operations. (File photo: iStock)

Consequences.

Consequences.

- How will the app handle private, sensitive medical data?
- Is the app secure and reliable enough?
- What if a false heart attack is detected?
- What if a heart attack is not detected?
- How could it impact healthcare costs, job markets and education?

Discussions around the topic.

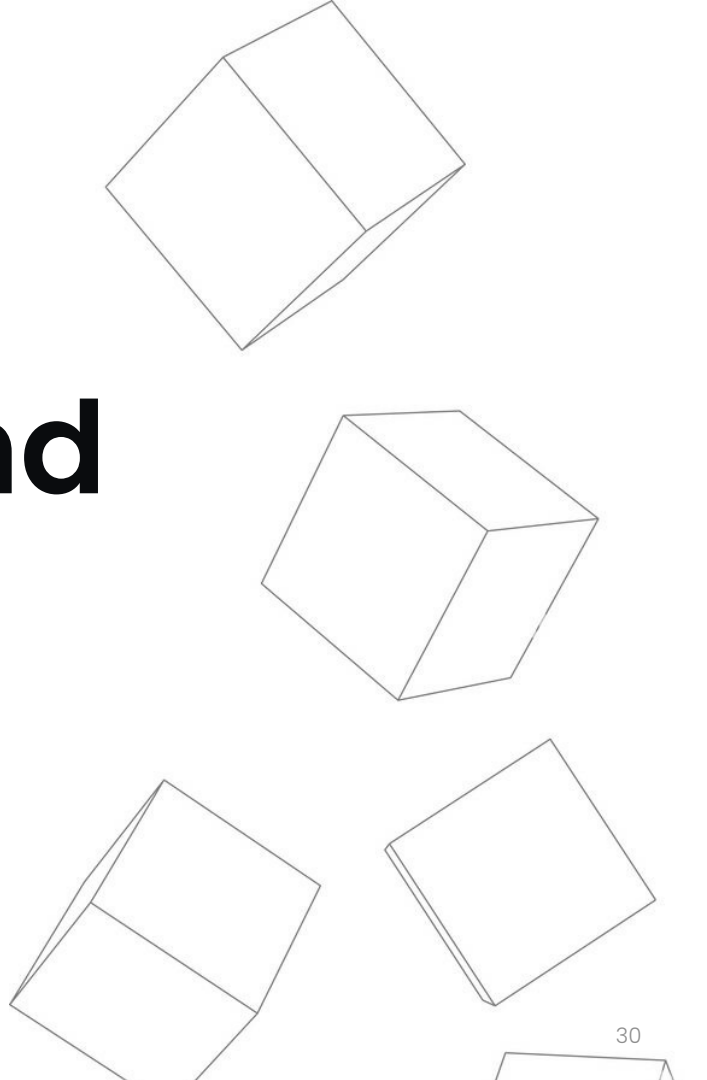
The issue of utilitarianism

“The greatest good for the greatest number”.

What do do.

- Implement a risk assessment framework (AI Risk Repository)
- Ensure that Ai is aligned with values, goals and norms
- Respect cultural and individual diversity
- Ensure inclusivity and that everyone have access to the benefits of AI

Transparency and explainability.



Concepts.

Transparency

Visibility and openness of an AI system's design, data, and operation. It involves disclosing detailed information about the AI's development process, data used for training, functioning mechanisms, and deployment process.

Explainability

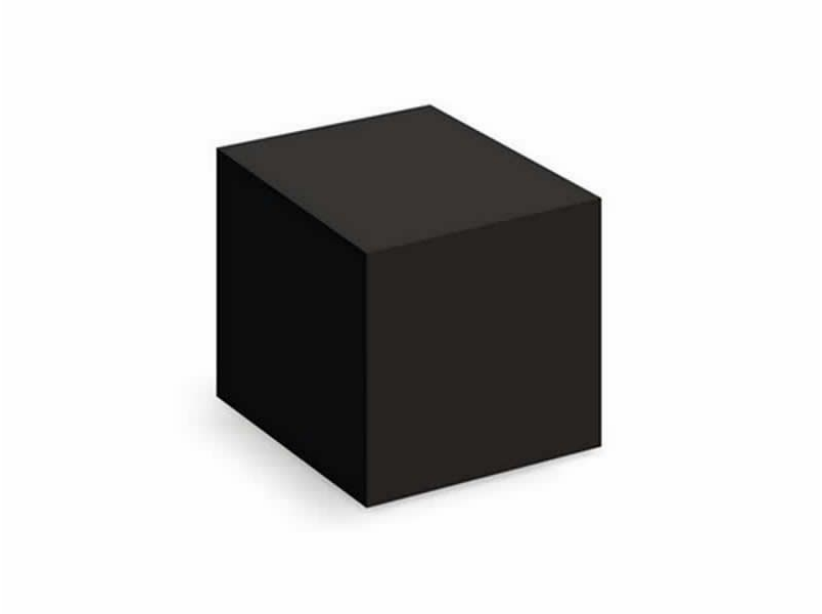
Focuses on making the complex decisions and outputs of an AI system understandable to users, regardless of their technical expertise.

What the principle says.

Transparency and explainability

AI should be designed for humans to easily perceive, detect, and understand its decision process. Transparency reassures us that AI systems operate ethically and responsibly.

The problem.



Why is it important.

- To justify the decision-making process
- Prevent user errors
- Users have the right to know how their data is used
- A moral obligation to understand the consequences of our actions
- To provide trust

How to make models more transparent?

- Use simpler models combined with more sophisticated models
- Track relevant dependencies between inputs and outputs
- Follow the latest research (XAI Based Intelligent Systems for Society 5.0)
- Clearly communicate to the user
  - Chances of errors
  - How personal data is being collected and used

Discussions around the topic.

Copyright issues

- Is AI infringing copyright laws by using art without consent from the artists?
- Can I copyright my AI generated artwork?

Discussions around the topic.

How much is too much?

The more that is revealed about the algorithms and the data, the more harm a malicious actor can cause.

Accountability and responsibility.

What the principle says.

Accountability and responsibility

AI systems should be auditable and traceable. There should be oversight, impact assessment, audit and due diligence mechanisms in place to avoid conflicts with human rights norms and threats to environmental wellbeing.

Who is responsible when AI fails?

- Heart attack preventive app makes a **wrong diagnose**.
- Autonomous vehicles make **wrong decision** and crashes a building.
- Identified **bias** against black people in a recruiting software
- Personal user **data leaked**
- ...

Who is responsible when AI fails?

It's complicated...

- Why did the error occur?
- AI systems are unpredictable
- Problem of "many hands"
- Are people more or less guilty depending on the amount of responsibility?
- What if the program was modified on the client-side?

Who is responsible when AI fails?

Consider the following two sentences:

1. “The AI system made a mistake.”
2. “The programmers/company made a mistake in the AI system.”

Consequences.

Tesla

Tesla Autopilot feature was involved in 13 fatal crashes, US regulator says

Federal transportation agency finds Tesla's claims about feature don't match their findings and opens second investigation

Guardian staff and agencies

Fri 26 Apr 2024 21.20 CEST

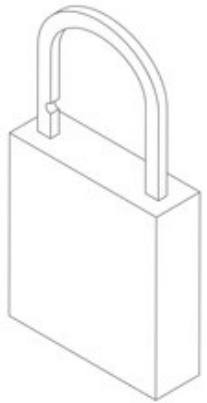
Share

US auto-safety regulators said on Friday that their investigation into Tesla's Autopilot had identified at least 13

What we can do.

- Transparency is key
- Access risk and oversee processes
- Use an accountability framework (AP4AI)
- Understand national and international laws, regulations, and guidelines that your AI may have to work within

Privacy and data protection.



What the principle says.

Privacy and data protection

Users and their data should be treated with dignity and respect. Privacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworks should also be established.

Discussions around the topic.

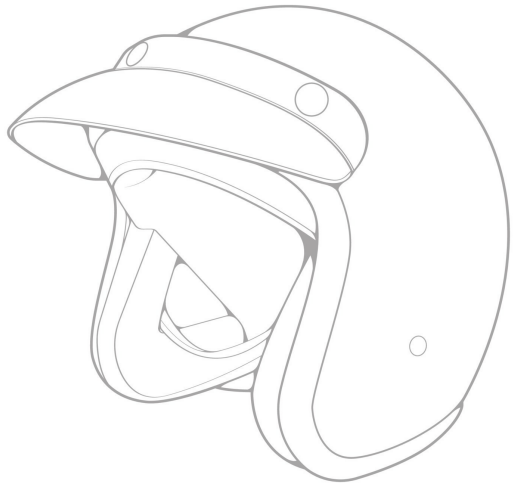
What exactly is your “own data”?

Is it the raw data, or the collected and analyzed data? If the data is used for secondary purposes, is it still your data?

What we can do.

- Follow General Data Protection Regulation (GDPR) guidelines
- Use data anonymization methods to protect privacy (Generalization, pseudonymization, Synthetic data)

Safety and security.



What the principle says.

Safety and security

Unwanted harms (**safety** risks) as well as vulnerabilities to attack (**security** risks) should be avoided and addressed by AI actors.

Concerns regarding safety.

Safety in AI

- Robustness
- Be safe and predictable

Strong testing

Ai as a threat

- Future threats to humanity
- Superintelligence

Human oversight

Discussions around the topic.

What are the acceptable limits to robustness?

There will be a set of circumstances so incredible that even if the system's safety cannot be assured. Where this limit is, though, is a difficult problem, and definitely not one that is exclusive to AI or even technology.

Discussions around the topic.

Can AI make the world safer?

Can AI make the world feel safer? Safer for whom? What if it clashes with other human rights?

Discussions around the topic.

This photograph shows a CCTV surveillance camera with a logo of the Paris 2024 Olympic and Paralympic Games displayed on the Grand Palais Olympic site in the background in Paris on July 22, 2024.
Photographer: Emmanuel Dunand/AFP via Getty Images

July 26, 2024, 9:30 AM GMT+1

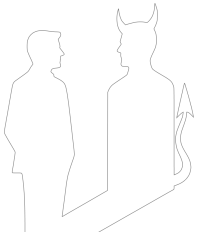
Olympics' AI Security Stokes Backlash Over Mass Surveillance

Jorja Siemons
Reporter

- Video security tool sparks larger discussion of AI in security
- World Cup, Olympics among mega-events soon coming to US

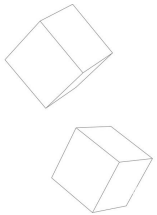
01

Fairness and non-discrimination



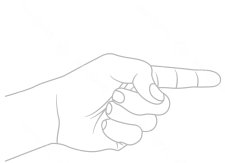
02

Non-maleficence



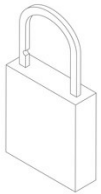
03

Transparency and explainability



04

Accountability and responsibility



05

Privacy and data protection

06

Safety and security

Final notes.

The issue of principles.

“[Ethics] plays the role of a bicycle brake on an intercontinental airplane”

Ulrich Beck, 1988

Our role.

Responsability

Act responsibly, be mindful about what we are creating and the possible consequences

Collaboration

Work with professionals and companies that are specialist in the subject, like ethicists or legal experts

Continuous learning

Encourage ongoing education and awareness in AI ethics

Advantages of an Ethical approach to AI.

01

Anticipate and avoid costs

02

Find opportunities that bring social value

Sustainable Development Goals.

References and further reading.

Ethics of AI online course:

- Ethics of AI - MOOC.fi

Google's guideline on responsible data collection:

- People + AI Guidebook

Bias in AI systems:

- Mitigating Bias in AI
- Towards a Standard for Identifying and Managing Bias in AI

Transparency and explainability:

- European data protection supervisor - Explainable AI
- XAI Based Intelligent Systems for Society 5.0
- Artificial intelligence and copyright: use of generative AI tools to develop new content

AI Risk assessment:

- AI Risk repository

Accountability:

- Accountability principles for AI
- Accountability in artificial intelligence: what it is and how it works

Ethics washing:

- From Ethics Washing to Ethics Bashing: A View on Tech Ethics from Within Moral Philosophy

Promoting AI for good:

- AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations
- The AI for People Podcast

ProductDock

Thank you