DEMOCRACY AT RISK: EU POLITICS IN THE 21ST CENTURY

UNIVERSITY of the PELOPONNESE

JEAN MONNET MODULE 2023-2025

ANASTASIOS GIANNAROS

TODAY'S DISCUSSION

UNIVERSITY of the PELOPONNESE

- What is AI?
- Content Moderation via AI
- Round table discussion: How can AI change our world?

Artificial Intelligence
what is it?

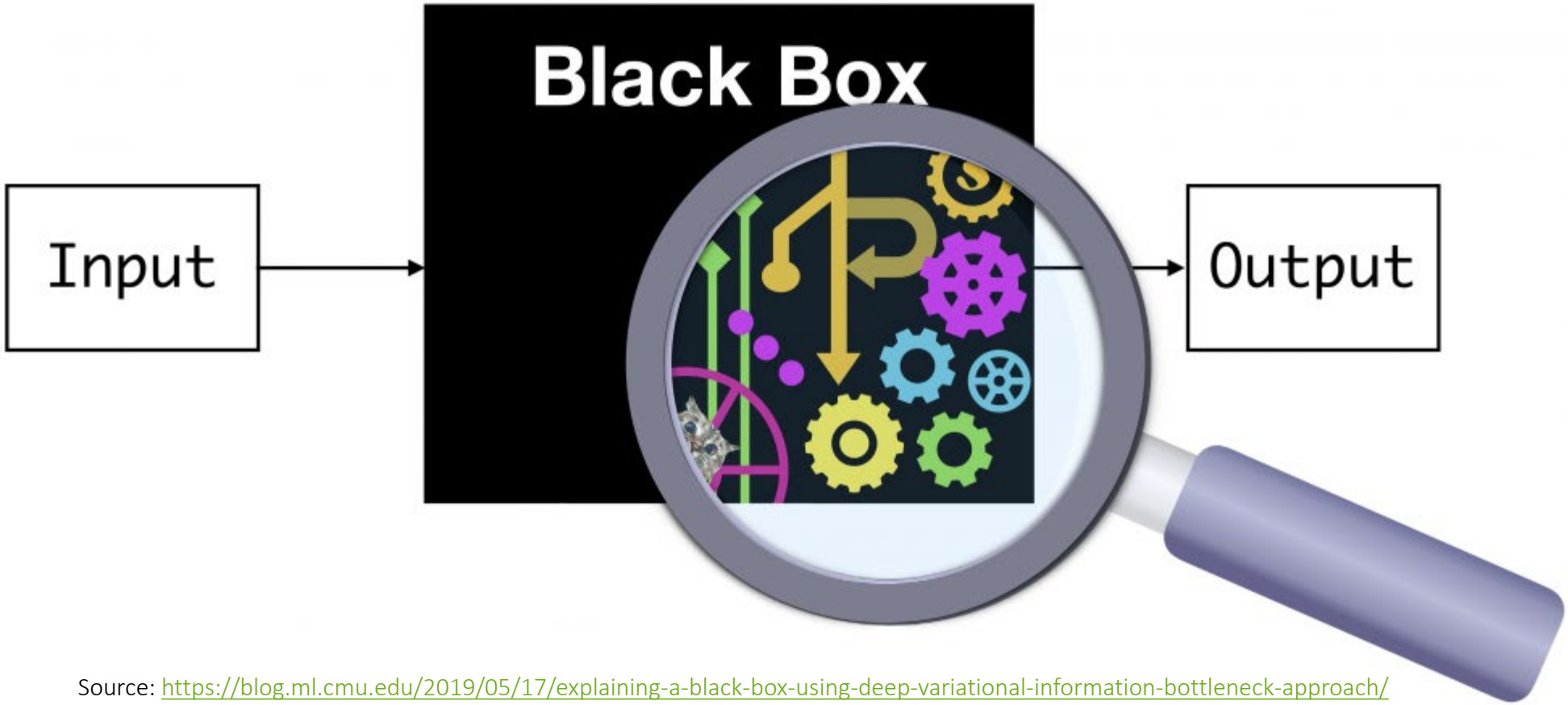
UNIVERSITY of the PELOPONNESE

Software that is trained on very large amount of data with the goal of carrying out a very specific task imitating humans

Artificial Intelligence
how is it trained?

Humans provide a problem and the solution to it. Considering there are enough such pairs (problems & solutions), the AI can learn how to recognize patterns, perform tasks and make decisions that will provide the desired/expected output.

Artificial Intelligence



Source: https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Artificial Intelligence
example #1

Virtual Personal Assistants

AI-powered virtual assistants like Siri, Google Assistant, and Alexa use natural language processing and machine learning algorithms to understand and respond to user queries and perform tasks such as setting reminders, making recommendations, and providing information.

Artificial Intelligence
example #2

Recommendation Systems

AI algorithms are used in recommendation systems employed by platforms like Netflix, Amazon, and Spotify to analyze user preferences and behavior, and provide personalized content and product recommendations.

Artificial Intelligence
example #3

Image and Speech Recognition

AI technologies are used in image and speech recognition systems, enabling applications such as facial recognition, object detection, speech-to-text conversion, and voice authentication.

Artificial Intelligence
example #4

Autonomous Vehicles

Self-driving cars utilize AI algorithms and sensors to navigate and make decisions in real-time, analyzing their environment and avoiding obstacles.

Artificial Intelligence
example #5

UNIVERSITY of the PELOPONNESE

Fraud Detection

AI is used in financial institutions to detect fraudulent activities by analyzing patterns, anomalies, and historical data, helping identify potential risks and prevent financial crimes.

Artificial Intelligence
example #6

Healthcare and Diagnostics

AI is employed in healthcare for tasks such as medical imaging analysis, disease diagnosis, drug discovery, and personalized medicine, enabling more accurate and efficient medical interventions.

Artificial Intelligence
example #7

UNIVERSITY of the PELOPONNESE

Natural Language Processing

AI-powered natural language processing (NLP) systems are used to analyze and understand human language, facilitating applications like chatbots, language translation, sentiment analysis, and text summarization.

Artificial Intelligence
example #8

Social Media Analysis

AI algorithms are employed by social media platforms to analyze user behavior, detect hate speech, identify fake accounts and inappropriate content, and enhance content moderation efforts.

Artificial Intelligence
example #9

Customer Service Automation

AI-powered chatbots and virtual assistants are used in customer service to automate responses, provide instant support, and handle routine inquiries, improving customer experiences.

Artificial Intelligence
example #10

Gaming

AI is utilized in gaming for tasks such as creating intelligent non-player characters (NPCs), generating realistic behaviors, and optimizing gameplay experiences.

Artificial Intelligence
example #11

Business Intelligence

AI is used in business intelligence systems to analyze large volumes of data, extract insights, and make data-driven decisions. AI algorithms can process complex data sets, identify patterns, and provide valuable business insights, helping organizations optimize operations, improve decision-making, and identify market trends.

Artificial Intelligence
example #12

Targeted Ads

AI is employed in targeted advertising and marketing campaigns to analyze user data, behavior, and preferences. AI algorithms can identify relevant audiences, segment users based on their interests and demographics, and deliver personalized advertisements and marketing messages, increasing the effectiveness and ROI of advertising efforts.

Artificial Intelligence
End User's Everyday Use

UNIVERSITY of the PELOPONNESE

- ChatGPT
- Midjourney
- DALL-E
- Deep Art

Artificial Intelligence
… in Content Moderation?

UNIVERSITY of the PELOPONNESE

Could AI enhance the content moderation procedure?

How?

- Pictures
- Text
- Speech

On Social Media?

Artificial Intelligence
Content Moderation

UNIVERSITY of the PELOPONNESE

Techniques in Content Moderation:

- Keyword Filtering
- Image Recognition
- Sentiment Analysis
- Anomaly Detection

Artificial Intelligence
Techniques in Content Moderation

UNIVERSITY of the PELOPONNESE

Keyword Filtering

Keyword filtering involves creating a list of specific words or phrases that are considered inappropriate or violate platform guidelines. The AI algorithm scans user-generated content and flags or filters out any instances that contain these predefined keywords.

Artificial Intelligence
Techniques in Content Moderation

UNIVERSITY of the PELOPONNESE

Image Recognition

AI algorithms are trained to analyze images and identify specific visual patterns or objects that may violate platform policies, such as explicit content or violent imagery.

Artificial Intelligence
Techniques in Content Moderation

UNIVERSITY of the PELOPONNESE

Sentiment Analysis

Sentiment analysis involves analyzing the emotional tone or sentiment expressed in user-generated content. AI algorithms can automatically assess whether the sentiment is positive, negative, or neutral.

Artificial Intelligence
Techniques in Content Moderation

UNIVERSITY of the PELOPONNESE

Anomaly Detection

Anomaly detection focuses on identifying unusual or abnormal patterns in user-generated content. AI algorithms learn from a large dataset of normal or expected behavior and can flag any deviations from this baseline.

Artificial Intelligence
… in Content Moderation?

UNIVERSITY of the PELOPONNESE

In March 2021, at one of many congressional hearings on Facebook, CEO Mark Zuckerberg reported:

"More than 95% of the hate speech that we take down is done by an AI and not by a person… And I think it's 98 or 99% of the terrorist content that we take down is identified by an AI and not a person."

Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation, 117th Cong. (2021) (statement of Mark Zuckerberg, Facebook), https://energycommerce.house.gov/committee-activity/hearings/hearing-on-disinformation-nation-social-medias-role-in-promoting.

Artificial Intelligence
Content Moderation

Content classification

Is content good or bad? Appropriate or inappropriate?

Artificial Intelligence
Controversial Topics Examples

Political Speech

What can be considered as acceptable political speech in one country could be viewed as controversial or even illegal in another. Issues such as hate speech, extremist ideologies, or inflammatory content can provoke debates about the boundaries of free speech and the responsibility of digital platforms to moderate such content.

Artificial Intelligence
Controversial Topics Examples

UNIVERSITY of the PELOPONNESE

Nudity and Sexual Content

Content related to nudity, sexual acts, or explicit material can be highly controversial and subject to varying cultural and legal standards. What is considered acceptable or even artistic in some contexts may be deemed inappropriate or offensive in others.

#freethenipple

Artificial Intelligence
Controversial Topics Examples

Violence and Graphic Content

The portrayal of violence, graphic imagery, or disturbing content can raise debates about the limits of acceptable content. Depictions of violence, including graphic scenes from real-life incidents or fictional portrayals, can be subject to differing cultural sensitivities and legal restrictions.

Artificial Intelligence
Controversial Topics Examples

Hate Speech and Discriminatory Content

Hate speech and discriminatory content that targets specific racial, ethnic, religious, or social groups can be highly contentious.

#MyBodyMyChoice

Artificial Intelligence
Controversial Topics Examples

UNIVERSITY of the PELOPONNESE

Misinformation and Disinformation

The spread of false information, conspiracy theories, and disinformation is a significant concern in the digital age. Determining what constitutes false or misleading information can be challenging, as it often involves complex judgments about intent, context, and potential harm.

Artificial Intelligence
… in Content Moderation?

UNIVERSITY of the PELOPONNESE

In March 2021, at one of many congressional hearings on Facebook, CEO Mark Zuckerberg reported:

"More than **95%** of the **hate speech** that we take down **is done by an AI** and not by a person… And I think it's **98 or 99%** of the **terrorist content** that we take down **is identified by an AI** and not a person."

Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation, 117th Cong. (2021) (statement of Mark Zuckerberg, Facebook), https://energycommerce.house.gov/committee-activity/hearings/hearing-on-disinformation-nation-social-medias-role-in-promoting.

Artificial Intelligence
… in Content Moderation?

UNIVERSITY of the PELOPONNESE

How Important is it to have efficient Content Moderation techniques?

So… Human or AI content moderation?

Artificial Intelligence
Human Content Moderation

Pros:

- Contextual Understanding: Humans can better understand the nuances, cultural references, and context behind content, allowing for more accurate judgment and decision-making.
- Flexibility and Adaptability: Human moderators can adapt quickly to changing trends, emerging issues, and new forms of problematic content.
- Subjective Judgment: Humans can bring subjective judgment and ethical considerations to the decision-making process, taking into account the intent, cultural differences, and potential impact of content.
- Empathy and Emotional Intelligence: Human moderators can empathize with users, understanding their concerns, and providing a more compassionate response when dealing with sensitive issues

Artificial Intelligence
Human Content Moderation

UNIVERSITY of the PELOPONNESE

Cons:

- Scale and Volume: With the enormous amount of content generated (on social media platforms), human moderators may struggle to keep up with the volume and pace of content moderation, leading to potential delays and backlogs.
- Bias and Subjectivity: Human moderators can bring their own biases and subjective interpretations to content moderation decisions, which may result in inconsistencies and potential unfairness.
- Cost and Scalability: Hiring and training a large team of human moderators can be costly and may limit scalability, especially for platforms with global user bases.
- Emotional Toll: Moderating disturbing or offensive content can take a toll on the mental well-being of human moderators, leading to issues such as burnout, compassion fatigue, and psychological stress.

Artificial Intelligence
AI Content Moderation

Pros:

- Scalability and Efficiency: AI algorithms can process vast amounts of content quickly, enabling efficient moderation and addressing the scale of user-generated content
- Consistency and Objectivity: AI systems can apply predefined rules and guidelines consistently without being influenced by personal biases or subjective interpretations.
- Speed and Real-Time Monitoring: AI algorithms can detect and flag potentially problematic content in real-time, allowing for quicker response and action.
- Reduced Human Exposure to Harmful Content: AI moderation can minimize the exposure of human moderators to harmful or disturbing content, reducing the potential negative impact on their well-being.

Artificial Intelligence
AI Content Moderation

Cons:

- Lack of Contextual Understanding: AI systems may struggle to fully understand the context, sarcasm, or subtle nuances of certain types of content, leading to potential false positives or false negatives in moderation decisions.
- Overreliance on Algorithms: Relying solely on AI moderation can lead to the risk of automated censorship, where legitimate content may be wrongly flagged or suppressed.
- Training Biases and Algorithmic Fairness: AI algorithms can inherit biases from training data, resulting in discriminatory or unfair moderation decisions, particularly towards marginalized communities.
- Constant Evolution of Problematic Content: AI algorithms may struggle to keep up with the ever-evolving nature of problematic content, such as new forms of hate speech or emerging trends, requiring continuous updates and improvements.

Artificial Intelligence
Content Moderation

UNIVERSITY of the PELOPONNESE

AI vs Human Content Moderation

Which one is better?

Artificial Intelligence
AI vs Human Content Moderation

UNIVERSITY of the PELOPONNESE

- Twitter: 500 million tweets per day
- Instagram: more than 50 million pictures so far

Source: Salman Aslam, “Instagram by the Numbers: Stats, Demographics & Fun Facts,” Omnicore, February 27, 2022, https://www.omnicoreagency.com/instagram-statistics.

Artificial Intelligence
Content Moderation

How easy is it to decide whether some content is legal/acceptable or illegal/unacceptable?

Determining (il)legality is a matter of applying subjective set of factors using good old-fashioned legal judgement, predictions about how a court might rule, and risk management. On that thought, have in mind that amongst different regions, the rules and the application of the law varies.

It is said that outside of the court, what is (il)legal is just a personal opinion.

Artificial Intelligence
Content Moderation

Source: Daphne Keller (*platform regulation director, Stanford Cyber Policy Center*)

Artificial Intelligence
Content Moderation

So, who makes the decision on social media platforms about the (il)legality of a post?

→ The company that has developed and is maintaining the platform.

Why?

→ Someone has to. It is their obligation in an effort to keep risks low and profits high. They do not want the users complaining and/or abandoning the platform.

Artificial Intelligence
Object Classification

What is Object Classification?

→ Object Classification is the process of identifying and assigning labels to specific objects.

How is Object Classification used in automated tasks?

→ Having trained an AI on large labeled dataset, it can then categorize/classify future objects by itself.

Artificial Intelligence
Object Classification

In the case of classifying whether an object in an image is a dog or an apple it is easy to train/teach the AI because most of the people will agree on the answer.

If the question though is whether some content contains hate speech, it is not that easy. Why? Because most people won't agree on the answer.

Artificial Intelligence
Object Classification

"Why is it so difficult to block spam/inappropriate content"

«Don't confuse a subjectivity problem for an accuracy problem, especially when you're using automation technology» - Alex Feerst (2022)

If humans can't agree on something, how do we expect different AIs to agree and come to a common conclusion?

Artificial Intelligence
Object Classification

UNIVERSITY of the PELOPONNESE

Over time, AI can become more efficient than humans. So, will it make decisions on its own? Who takes responsibility for these?

Artificial Intelligence
AI Content Moderation

UNIVERSITY of the PELOPONNESE

The endless discussion does not help in the case of social media at least. A decision has to be made about the content at hand and usually immediately. After all, behind every social media there is a company with all that entails.

Artificial Intelligence
AI Content Moderation

UNIVERSITY of the PELOPONNESE

How much does it cost to have human moderators on social (media) platforms?

- How many people?
- Who are they? Background.
- From/On which country?
- Employment status (*full/part time, contractual, freelance, interns, volunteer*)

Artificial Intelligence
… in Content Moderation?

UNIVERSITY of the PELOPONNESE

In March 2021, at one of many congressional hearings on Facebook, CEO Mark Zuckerberg reported:

"More than 95% of the hate speech that we take down is done by an AI and not by a person… And I think it's 98 or 99% of the terrorist content that we take down is **identified** by an AI and not a person."

Is identification another word for discrimination?

Disinformation Nation: Social Media's Role in Promoting Extremism and Misinformation, 117th Cong. (2021) (statement of Mark Zuckerberg, Facebook), https://energycommerce.house.gov/committee-activity/hearings/hearing-on-disinformation-nation-social-medias-role-in-promoting.

Artificial Intelligence
AI Content Moderation

UNIVERSITY of the PELOPONNESE

How should AI react when it detects inappropriate content with a certainty above a certain percentage (e.g. 80%)?

→ Auto block?
→ Does it send it to a human for a final decision?
→ Monitor the reactions to this content and act accordingly?

Artificial Intelligence
AI Content Moderation

UNIVERSITY of the PELOPONNESE

It is not only the content itself that needs moderation by AI.
Disinformation campaign (disinformation/propaganda).

In this case, what are we looking for? **Metadata!**

→ Which accounts post content and/or interact with it
→ Who lies behind the account in question (IP address, patterns in actions,…)
→ Whether they had activity in previous campaigns
→ When did the account go live and/or start being active
→ Whether the account's profile image is copied from somewhere else
→ Whether the account is associated with suspicious groups

Artificial Intelligence
AI Content Moderation

Are people OK with their content being moderated by a machine?

Suggestions?

AI Transparency

Artificial Intelligence
AI Content Moderation

UNIVERSITY of the PELOPONNESE

Transparency in AI

- Why did AI take this decision? AI should be able to inform if asked. EU Law.
- Is it always easy though to answer this?
- Even if it is, why should we trust the answer?
- Shouldn’t the AI’s algorithm be made publicly available? Open Source? Or should it?

Artificial Intelligence
Final Considerations

- Harmful users can also utilize AI systems to tackle AI content moderation. Who will win in the end? This whole thing is like the cat and mouse game.
- Hackers are now using AI to hack systems.

Artificial Intelligence
Final Considerations

UNIVERSITY of the PELOPONNESE

- AI content production? Intellectual Property Rights on the produced content?
- Will we reach a point in the future where all content (news, images, courses) will be generated by AI?
- If so, what implications will this have for humanity? Will the internet and our world itself be dominated by AI?

Artificial Intelligence
Final Considerations

UNIVERSITY of the PELOPONNESE

Ethical Considerations

- AI-based content moderation raises ethical considerations related to privacy, freedom of expression, transparency, and accountability.
- There is a need to strike a balance between maintaining a safe and inclusive online environment while respecting users' privacy and freedom of speech.
- Transparency in the content moderation process, clear guidelines and appeals mechanisms, and ongoing monitoring and evaluation of AI models are essential to address ethical concerns.