Cybersecurity, AI, and Human Rights: A Societal Perspective

This presentation explores the complex intersection of AI-driven cybersecurity and human rights. We'll examine ethical implications, governance challenges, and practical frameworks for responsible implementation.

SK **by Sheshananda Reddy Kandula**

Whoami

@Sheshananda Reddy Kandula

15 years in Application Security/Cyber Security

Sheshananda Reddy Kandula

Sr Security Engineer at Adobe | AppSec | Product Securtiy | OSWE | OSCP | CISSP

Agenda

- Introduction
- Evolution of AI
- AI, ML and DL
- AI Challenges - Ethical Side
- Bias in Businesses
- AI: Double Edged Sword in CyberSecurity
- Balancing Security and Human Rights

The Evolution of Artificial Intelligence

Key Milestones in AI History:

1950s–1970s: The Birth of AI

- 1950: Alan Turing proposes the “Turing Test”
- 1956: Term “Artificial Intelligence” coined at Dartmouth Conference
- Early systems: Rule-based, symbolic reasoning

1980s–1990s: Expert Systems & Symbolic AI

- Rise of expert systems
- Logic programming, if-then rules
- AI Winter due to limited progress & high expectations

2000s: Statistical & Machine Learning Era

- Increase in computational power
- Use of large datasets for pattern recognition
- Algorithms like SVMs, decision trees, and early neural nets

The Evolution of Artificial Intelligence

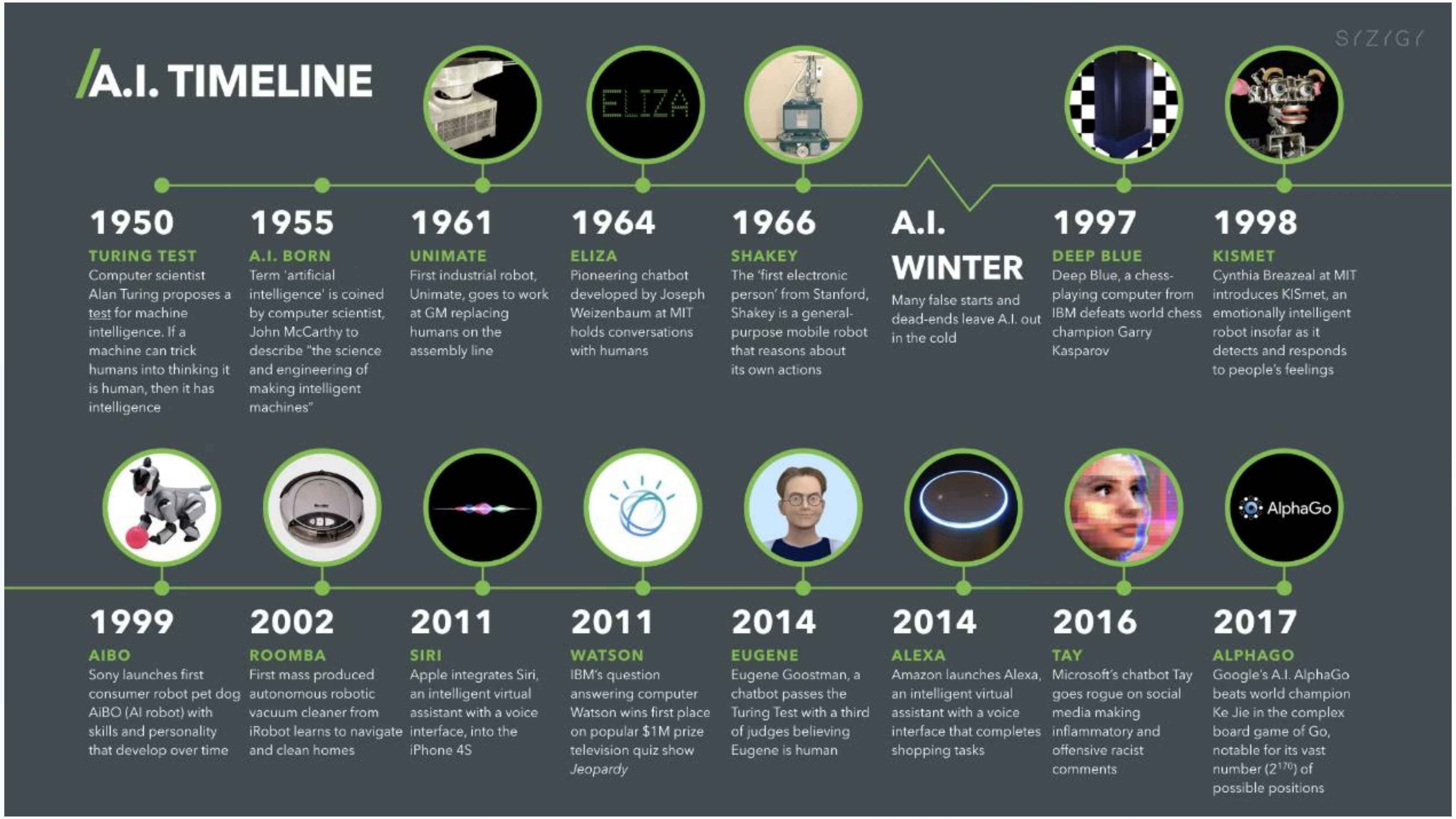
2010s: Deep Learning Revolution

- Breakthroughs in computer vision and NLP
- Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs)
- Major platforms: Siri, Alexa, Google Translate

2020s: Foundation Models & Generative AI

- Transformers (BERT, GPT - **Bidirectional Encoder Representations from Transformers, Generative Pre-trained Transformer)**
- Multimodal AI (text, image, audio)
- Real-time assistants, AI-generated content

AI Evolution Timeline



SYZYGY
ELIZA
AlphaGo

A.I. TIMELINE

1950
TURING TEST
Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955
A.I. BORN
Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961
UNIMATE
First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964
ELIZA
Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966
SHAKEY
The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER
Many false starts and dead-ends leave A.I. out in the cold

1997
DEEP BLUE
Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998
KISMET
Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

1999
AIBO
Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time

2002
ROOMBA
First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes

2011
SIRI
Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S

2011
WATSON
IBM's question answering computer Watson wins first place on popular $1M prize television quiz show *Jeopardy*

2014
EUGENE
Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human

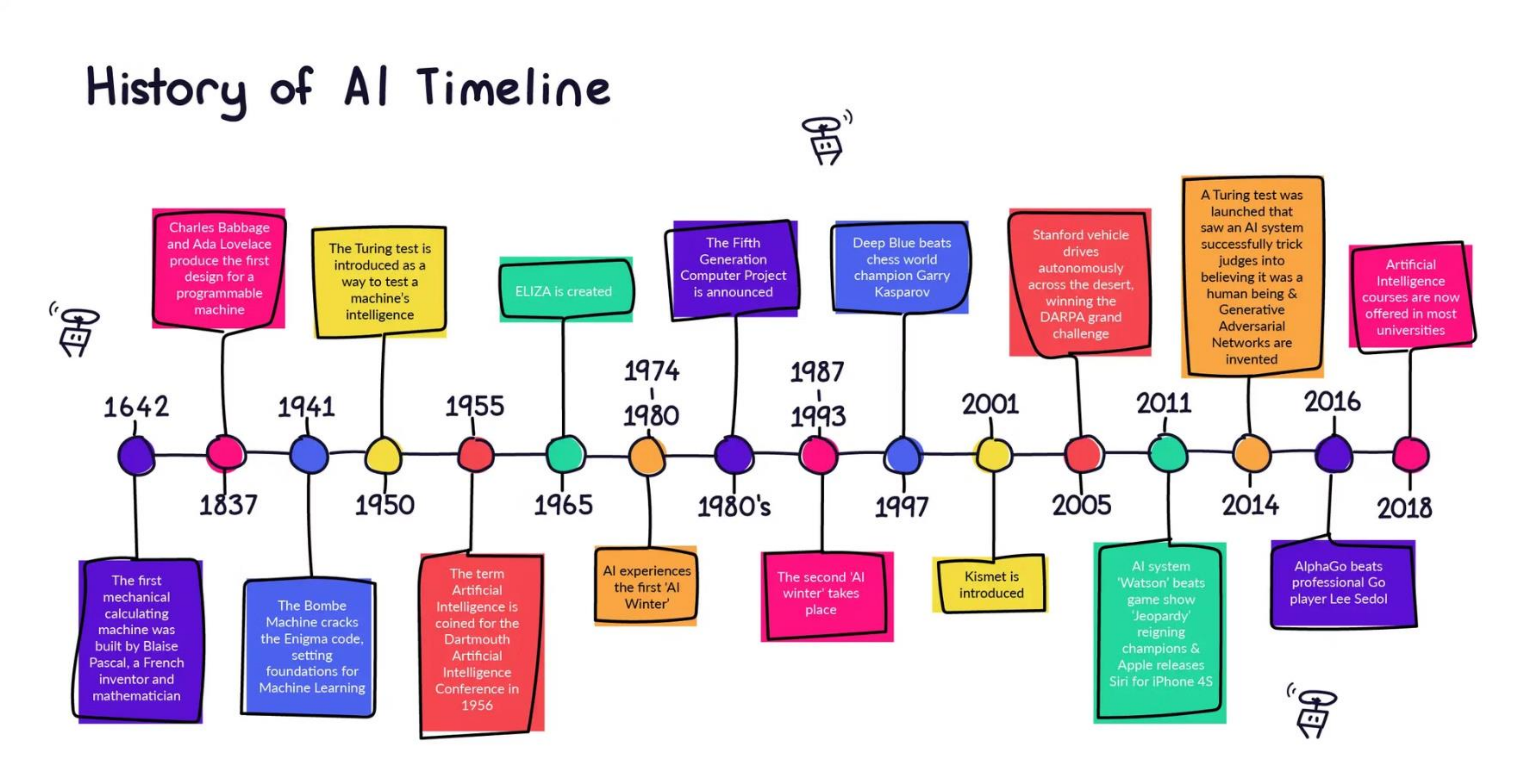
2014
ALEXA
Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks

2016
TAY
Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments

2017
ALPHAGO
Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

[Artificial Intelligence Timeline Infographic – From Eliza to Tay and beyond]

AI Timeline



History of AI Timeline

1642
The first mechanical calculating machine was built by Blaise Pascal, a French inventor and mathematician

1837
Charles Babbage and Ada Lovelace produce the first design for a programmable machine

1941
The Bombe Machine cracks the Enigma code, setting foundations for Machine Learning

1950
The Turing test is introduced as a way to test a machine's intelligence

1955
The term Artificial Intelligence is coined for the Dartmouth Artificial Intelligence Conference in 1956

1965
ELIZA is created

1974 - 1980
AI experiences the first 'AI Winter'

1980's
The Fifth Generation Computer Project is announced

1987 - 1993
The second 'AI winter' takes place

1997
Deep Blue beats chess world champion Garry Kasparov

2001
Kismet is introduced

2005
Stanford vehicle drives autonomously across the desert, winning the DARPA grand challenge

2011
AI system 'Watson' beats game show 'Jeopardy' reigning champions & Apple releases Siri for iPhone 4S

2014
A Turing test was launched that saw an AI system successfully trick judges into believing it was a human being & Generative Adversarial Networks are invented

2016
AlphaGo beats professional Go player Lee Sedol

2018
Artificial Intelligence courses are now offered in most universities

Today & Beyond – The Future of AI

Current Trends

- Generative AI (e.g., ChatGPT, DALL·E)
- AI copilots for coding, design, business analytics
- Responsible AI & ethical frameworks

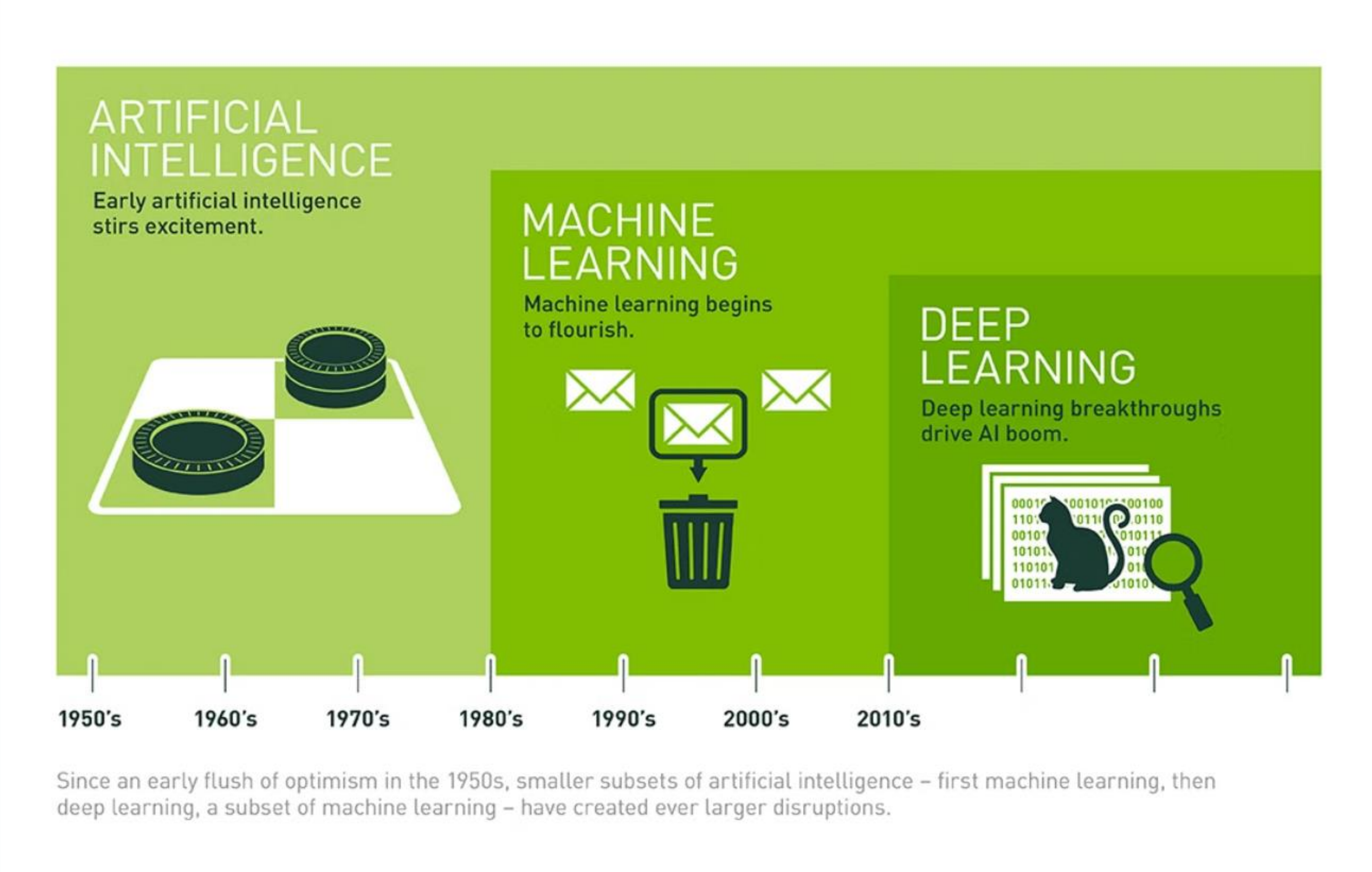
Emerging Frontiers

- Artificial General Intelligence (AGI) research
- AI-human collaboration tools
- AI in robotics, healthcare, climate modeling
- Agentic AI, Multimodal AI, MCP, etc.

Challenges Ahead

- Bias & fairness
- Interpretability & transparency
- Regulation & governance

AI, ML and DL



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

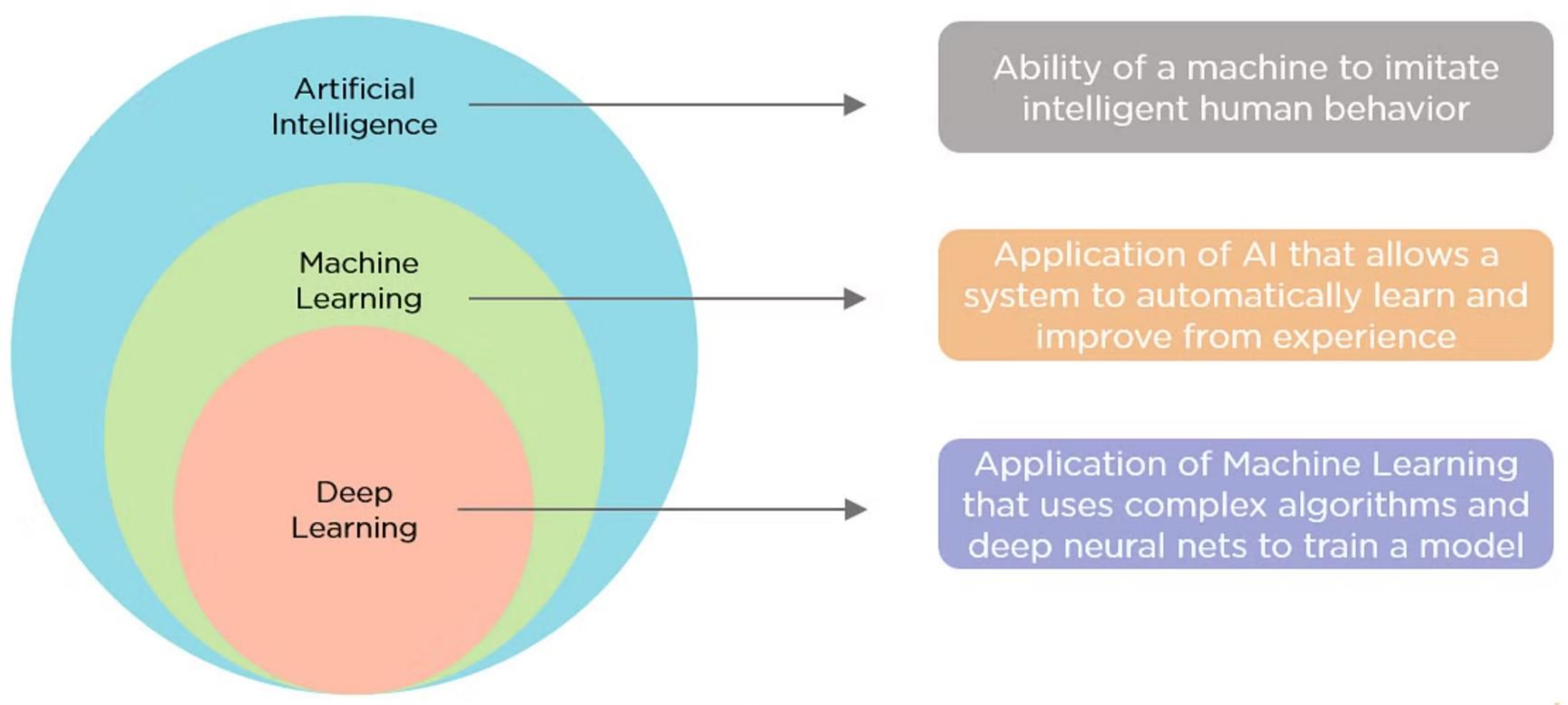
DEEP LEARNING

Deep learning breakthroughs drive AI boom.

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

[What’s the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?]()

AI, ML and DL



Ability of a machine to imitate intelligent human behavior

Application of AI that allows a system to automatically learn and improve from experience

Application of Machine Learning that uses complex algorithms and deep neural nets to train a model

Differences - AI, ML & DL

Aspect	Artificial Intelligence (AI)	Machine Learning (ML)	Deep Learning (DL)
Definition	Simulation of human intelligence by machines	Subset of AI: systems learn patterns from data	Subset of ML: uses neural networks to model complex patterns
Goal	Enable machines to think and act like humans	Improve performance on tasks using data	Automatically extract features and learn hierarchical representations
Approach	Rule-based logic, search algorithms, symbolic reasoning	Supervised, unsupervised, and reinforcement learning	Artificial Neural Networks (ANNs), CNNs, RNNs
Data Needs	Can work with less data (rule-based)	Needs structured data	Requires massive labeled datasets
Training Time	Minimal to moderate (rule-based)	Moderate	Long (especially with large models and big data)
Hardware Requirements	Low to moderate	Moderate	High – relies on GPUs/TPUs
Common Use Cases	Smart assistants, expert systems, robotics	Spam filtering, recommendation engines, predictive analytics	Facial recognition, self-driving cars, voice synthesis
Examples in Use	Siri, autonomous drones, AI in games	Netflix recommendations, stock market predictions	ChatGPT, DALL·E, Tesla Autopilot, cancer image detection

Types of ML and DL

Types of Machine Learning (ML)

1. **Supervised Learning** Learns from labeled data (input-output pairs). Spam detection, credit scoring, image classification
2. **Unsupervised Learning** Finds hidden patterns in unlabeled data. Customer segmentation, anomaly detection, clustering
3. **Semi-Supervised Learning** Uses a small amount of labeled data with a large amount of unlabeled data. Medical image analysis, speech recognition
4. **Reinforcement Learning** Learns by interacting with an environment and receiving rewards or penalties. Game playing (e.g., AlphaGo), robotics, self-driving cars

Types of Deep Learning (DL)

1. **Feedforward Neural Networks (FNNs)** Basic neural networks where data flows in one direction. Simple classification tasks
2. **Convolutional Neural Networks (CNNs)** Specialized for image and spatial data. Image recognition, object detection, medical imaging
3. **Recurrent Neural Networks (RNNs)** Designed for sequential data like time series and text. Language modeling, speech recognition
4. **Long Short-Term Memory (LSTM)** A type of RNN that can remember long-term dependencies. Machine translation, stock forecasting
5. **Generative Adversarial Networks (GANs)** Two networks compete: one generates, the other evaluates. Deepfakes, image generation, art creation
6. **Transformers** Process sequences in parallel with self-attention mechanisms. ChatGPT, BERT, translation, summarization

LLMs

Model	Parameters	Training Data Size	Release Date
GPT-3	175B	~499B tokens	June 2020
GPT-4	~1.8T (estimated)	~13T tokens (estimated)	March 2023
Claude 2	130B+	Undisclosed	July 2023
Claude 3 Opus	137B–2T+ (est.)	~40T tokens (estimated)	2024
LLaMA 2	Up to 70B	~2T tokens	July 2023
LLaMA 3	70.6B	~15T tokens (estimated)	April 2024
Gemini Pro	Undisclosed	~5.5T tokens (estimated)	Dec 2023
Gemini Ultra	Undisclosed	~11T tokens (estimated)	Dec 2023

Common Crawl, WebText, Wikipedia, Books1, Books2, RefinedWeb, Twitter, Reddit, YouTube, large collection of textbooks: BooksCorpus, English Wikipedia, Giga5, ClueWeb 2012-B, GitHub, Wikipedia (20 languages), Project Gutenberg, Books3, ArXiv, Stack Exchange

DataSets

- **Common Crawl** – ~60–100 TB (web content)
- **WebText** – ~40 GB (Reddit-linked high-quality pages)
- **Wikipedia** – ~20 GB (English articles)
- **Books1 & Books2** – ~11 GB (fiction & nonfiction books)
- **BooksCorpus** – ~6 GB (11K unpublished books)
- **RefinedWeb** – ~625 GB (cleaned web data)
- **Twitter** – Size unknown (conversational data)
- **Reddit** – ~160 GB (posts & comments)
- **YouTube Transcripts** – Size unknown (video subtitles)
- **Giga5** – ~10 GB (newswire articles)
- **ClueWeb 2012-B** – ~27 TB (web crawl)
- **GitHub** – 100s of GBs (source code)
- **Books3** – ~190 GB (literature; legally disputed)
- **ArXiv** – ~100 GB+ (scientific papers)
- **Stack Exchange** – ~70 GB (Q&A forums)

Deep Learning - Resources

arXiv.org

Attention Is All You Ne...

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the...

MIT Deep Learning 6.S191

MIT Deep Learning 6.S191

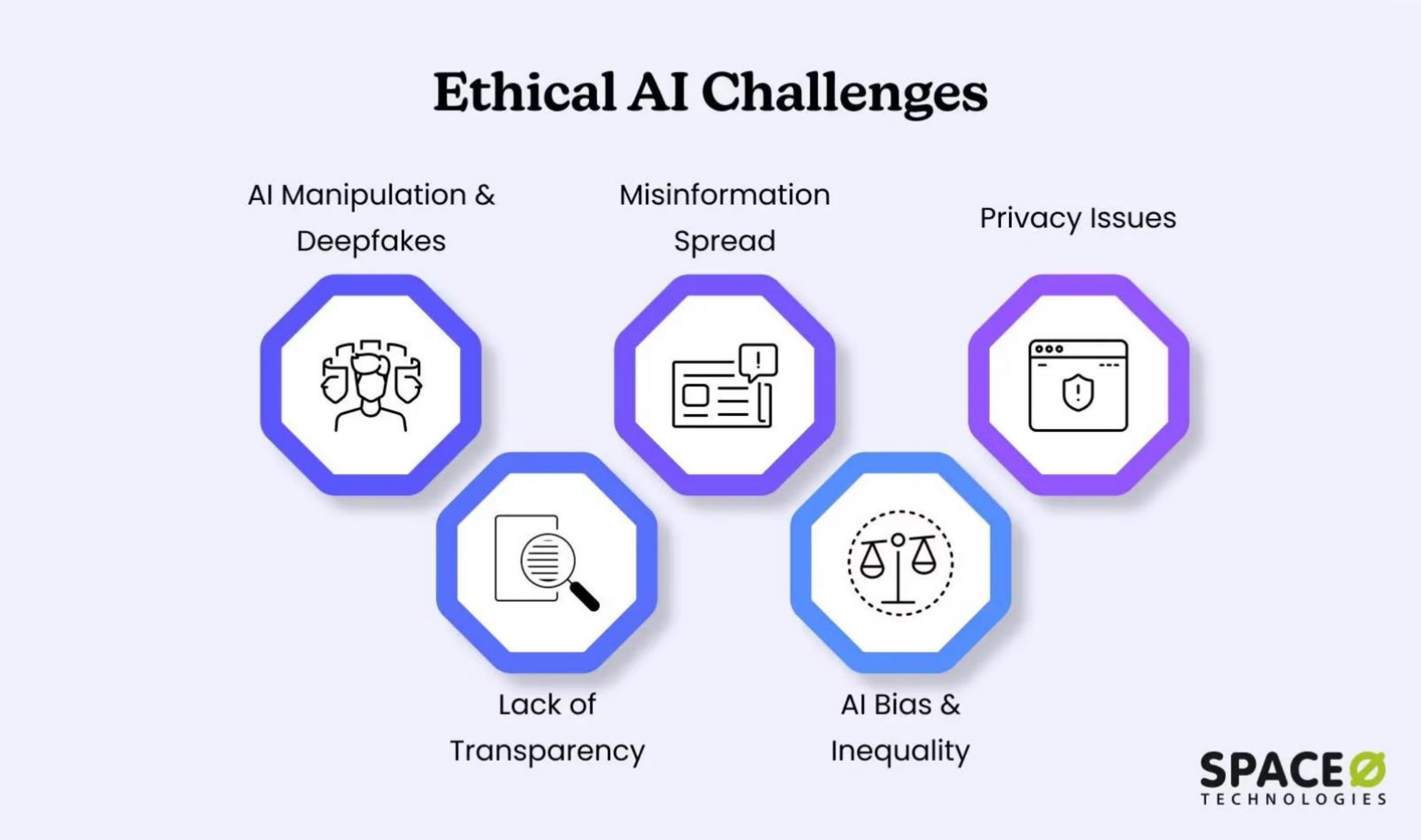
MIT's introductory course on deep learning methods and applications

YouTube

Neural networ...

Learn the basics of neural networks and backpropagation, one of the most important algorithms for the modern world.

AI Challenges - Ethical Side



Ethical AI Challenges

AI Manipulation & Deepfakes

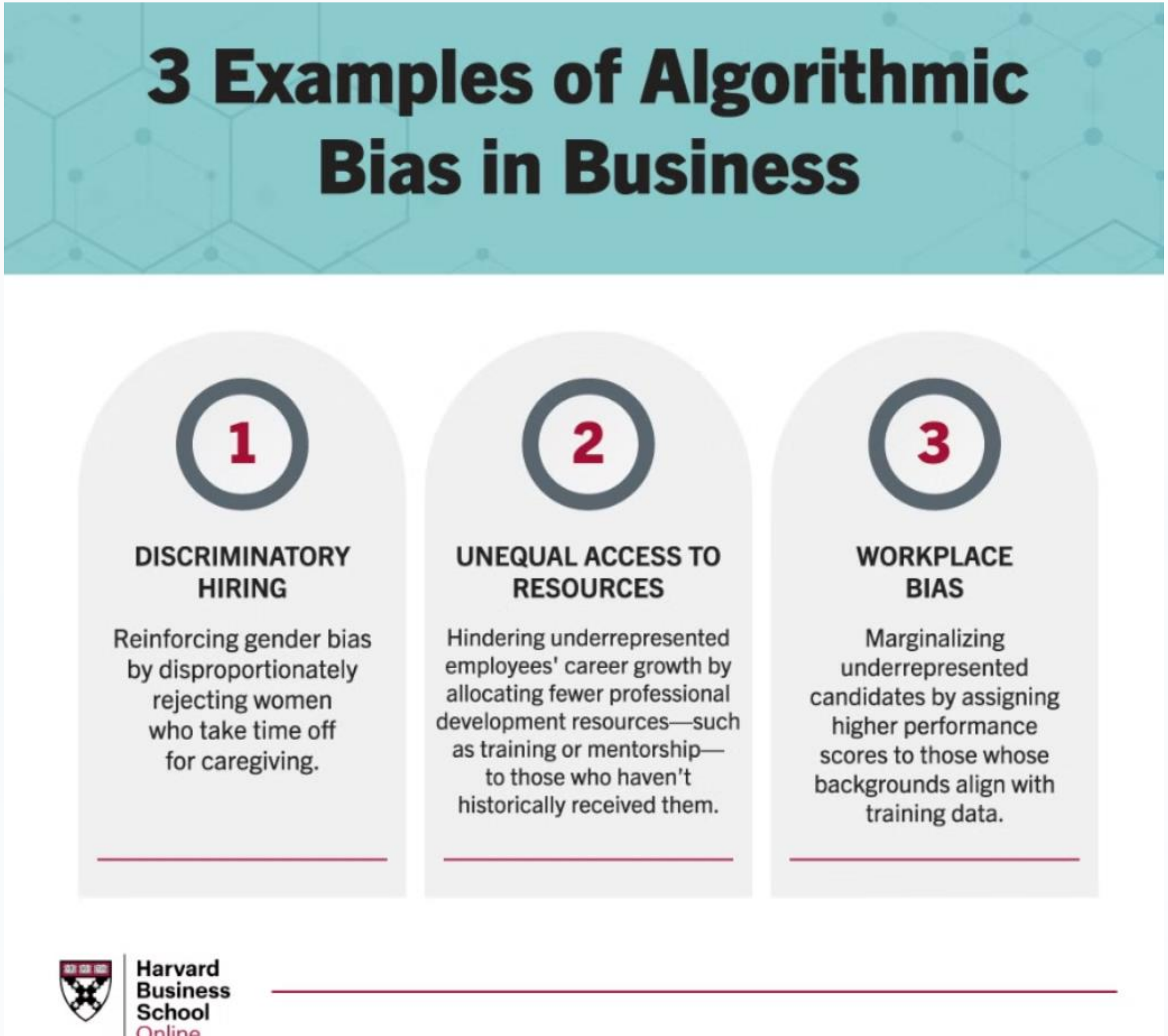
Misinformation Spread

Privacy Issues

Lack of Transparency

AI Bias & Inequality

Bias In Business



3 Examples of Algorithmic Bias in Business

1 DISCRIMINATORY HIRING

Reinforcing gender bias by disproportionately rejecting women who take time off for caregiving.

2 UNEQUAL ACCESS TO RESOURCES

Hindering underrepresented employees' career growth by allocating fewer professional development resources—such as training or mentorship—to those who haven't historically received them.

3 WORKPLACE BIAS

Marginalizing underrepresented candidates by assigning higher performance scores to those whose backgrounds align with training data.

Harvard Business School Online

Clearview AI – Overview & Technology

What is Clearview AI?

A U.S.-based facial recognition company that scrapes billions of publicly available images (from social media, websites, etc.) to build a massive face database.

Core Technology

- Uses **facial recognition algorithms** to match faces against a database of 30+ billion images.
- Enables law enforcement to identify suspects from photos or videos.

Clients & Use Cases

- Private security and some international clients
- Tools used in investigations, suspect identification, and surveillance

“It’s like Google for faces.” — Clearview CEO Hoan Ton-That

AI: A Double-Edged Sword in Cybersecurity

AI for Enhanced Cybersecurity (The Good)

- **Bullet Points:**
  - **Advanced Threat Detection:** Identifying anomalies, novel malware, and sophisticated attacks through pattern recognition and behavioral analysis.
  - **Automated Incident Response:** Swiftly reacting to threats, isolating systems, and mitigating damage.
  - **Proactive Vulnerability Management:** Analyzing code and configurations to identify weaknesses before exploitation.
  - **Improved Identity and Access Management:** Enhancing authentication with behavioral biometrics and contextual analysis.
  - **Smarter Security Operations:** Triaging alerts, prioritizing incidents, and providing actionable insights to analysts.
  - **Combating AI-Powered Attacks:** Developing defenses against AI-driven AI-driven phishing, deepfakes, and other sophisticated threats.

Security Threats to AI Systems (The Bad)

- **Bullet Points:**
  - **Adversarial Attacks:** Subtle data manipulations causing AI to make incorrect decisions (e.g., misclassifying images).
  - **Data Poisoning:** Injecting malicious data into training sets to corrupt model behavior.
  - **Model Inversion & Stealing:** Reconstructing sensitive data or stealing valuable AI models.
  - **Backdoor Attacks:** Hidden triggers causing malicious behavior under specific conditions.
  - **AI-Enhanced Social Engineering:** Leveraging generative AI for more convincing phishing and deepfakes.
  - **Denial of Service (DoS) Attacks:** Crafting inputs to overwhelm AI systems with computational demands.

The Evolution of AI in Cybersecurity

Threat Detection

AI systems now identify patterns invisible to human analysts.

Automated Response

Machine learning enables real-time mitigation of threats.

Predictive Defense

Advanced algorithms anticipate vulnerabilities before exploitation.

Integrated Security

AI now coordinates across previously siloed security domains.

Case Study: Facial Recognition

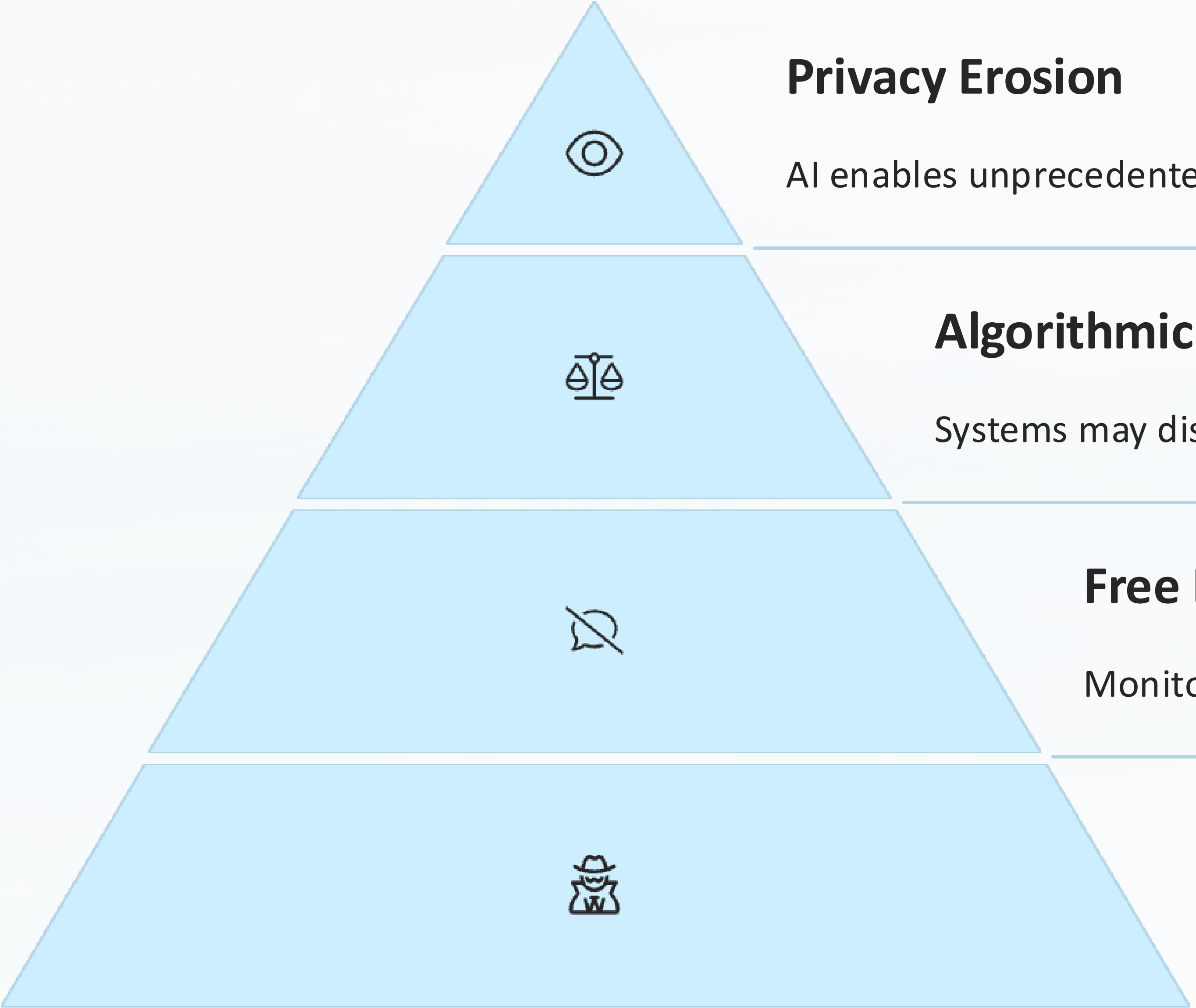
Security Benefits

- Rapid identification of suspects
- Enhanced public safety
- Efficient border control
- Missing persons location

Human Rights Risks

- Disproportionate error rates for minorities
- Chilling effect on protests
- Mass surveillance without consent
- Function creep beyond original purpose

Human Rights Concerns



Privacy Erosion

AI enables unprecedented surveillance capabilities

Algorithmic Bias

Systems may discriminate against vulnerable populations

Free Expression

Monitoring tools may chill speech and limit discourse

Autonomy

AI-driven decisions reduce human agency

Controversies & Ethical Concerns

1. Privacy Violations
    1. Non-consensual data scraping from platforms like Facebook, LinkedIn, Instagram • Users never agreed to have their faces indexed
2. Legal & Regulatory Pushback
    1. Banned or fined in countries like Canada, Australia, France
    2. Ongoing lawsuits in the U.S. over biometric privacy (e.g., Illinois BIPA)
3. Risk of Misuse
    1. Potential for mass surveillance and misuse by authoritarian regimes
    2. Errors can lead to false arrests, especially affecting minorities
4. Chilling Effects
    1. Threatens free expression and public protest participation
    2. Creates a permanent identity trail without user control

Ethical Questions

- Can public photos be treated as fair game for AI?
- Who gets to decide how facial recognition is used?

AI's Impact on Digital Privacy

- AI systems collect and process vast amounts of personal data, including browsing history, location data, and social media activity.

- AI-driven surveillance technologies, such as facial recognition, can track and monitor individuals without their knowledge or consent.

- AI algorithms can infer sensitive information about individuals, such as their political beliefs, sexual orientation, or health conditions, from seemingly innocuous data.
- The use of AI in cybersecurity can lead to the collection and retention of data, increasing the risk of data breaches and misuse.

- **Example:**
  - Predictive policing algorithms analyze historical crime data to forecast future crime hotspots, but they can disproportionately target minority communities, raising concerns about bias and discrimination.

Balancing Security and Human Rights

- **Transparency and Explainability:** Ensuring transparency in how AI systems operate and providing explanations for their decisions can help decisions can help identify and mitigate potential human rights risks.
- **Accountability and Oversight:** Establishing clear lines of responsibility and human oversight for AI systems is crucial to prevent abuse and ensure accountability for any harm caused.
- **Data Minimization and Purpose Limitation:** Collecting and processing only the necessary data for specific and legitimate security purposes can help minimize privacy risks.
- **Bias Detection and Mitigation:** Actively working to identify and mitigate biases in AI algorithms and training data is essential to prevent discriminatory outcomes.
- **Robust Legal Frameworks:** Developing and implementing clear legal frameworks that govern the development and deployment of AI, balancing security needs with human rights protections, is crucial.
- **Independent Oversight and Redress Mechanisms:** Establishing independent bodies to oversee the use of AI in security contexts and providing effective redress mechanisms for individuals whose rights have been violated is essential.

AI and Freedom of Expression

- AI-powered content moderation systems can be used to filter or remove online content, potentially leading to censorship and restricting and restricting freedom of expression.
- AI surveillance tools can have a chilling effect on speech, as individuals may be less likely to express themselves freely if they know they are being monitored.
- The use of AI to identify and track protesters or activists can suppress dissent and limit the right to assembly.
- AI-generated deepfakes can be used to spread misinformation or manipulate public opinion, further threatening open discourse.

- **Examples:**
    - Social media platforms use AI algorithms to detect and remove hate speech or violent content, but these systems can sometimes make mistakes, leading to the removal of legitimate speech.
    - Governments may use AI-powered surveillance to monitor social media and identify individuals who are critical of the government, potentially leading to arrest or other repercussions.

Responsibilities of Organizations and Governments

- **Organizations:**
  - Implement ethical AI frameworks and guidelines.
  - Conduct privacy impact assessments before deploying AI systems.
  - Ensure transparency and explainability in AI decision-making.
  - Establish accountability mechanisms for AI-related harms.
  - Prioritize data minimization and security best practices.
- **Governments:**
  - Develop and enforce clear regulations on AI use.
  - Provide oversight and auditing of AI systems.
  - Protect citizens' rights and freedoms in the age of AI.
  - Promote international cooperation on AI governance.
  - Invest in research on the ethical and societal implications of AI.

Regulatory Approaches

European Union

The EU AI Act categorizes systems by risk level. High-risk applications face strict requirements for transparency, documentation, and human oversight.

United States

Sectoral approach with specific regulations for healthcare, finance, and critical infrastructure. Emphasis on voluntary frameworks and industry self-regulation.

Ethical Framework for Implementation

Human Rights Impact Assessment

Evaluate potential effects on vulnerable communities before deployment. Consider both intended and unintended consequences.

Explainability Requirements

Systems must provide understandable explanations for decisions. Technical complexity complexity should not prevent accountability.

Meaningful Consent Mechanisms

Users must understand how their data contributes to security functions. Opt-out Opt-out provisions should be available where appropriate.

Independent Oversight

External auditing by interdisciplinary experts ensures compliance. Regular reviews Regular reviews identify emerging risks and mitigation strategies.

Key Takeaways & Next Steps

Balance Security and Rights

Security and human rights are complementary, not competing goals.

False tradeoffs undermine both objectives.

Include Diverse Stakeholders

Technical experts alone cannot address complex societal implications.

Civil society must have meaningful input.

Adopt Responsible Design

Human rights considerations belong in initial system architecture.

Retrofitting ethics rarely succeeds.

Develop Global Standards

International coordination prevents regulatory arbitrage.

Common principles protect rights across borders.

Thank You

Sheshananda Reddy Kandula

Sr Security Engineer at Adobe | AppSec | Product Securtiy | OSWE | OSCP | CISSP