AFMUN
UNHLAB-AI

UNDER SECRETARY GENERAL
Kuzey Karlık

UNDER SECRETARY GENERAL
Baran İnce

Type:
ANGER

Anime Series

01

UNHLAB AI GUIDE ↓

1. Letter From the Secretary General

Distinguished Delegates,

It is a great pleasure to welcome you all to AFMUN'26. As the Secretaries-General, we are truly pleased to see this conference once again bring together so many bright and motivated young individuals who believe in dialogue, diplomacy, and cooperation.

First and foremost, we would like to extend our sincere thanks to our academic and organization teams. This conference would not have been possible without their dedication and commitment, and we are deeply grateful for the effort they have put in.

We are living in a time where it is almost impossible to ignore what is happening around the world. Many of us often find ourselves thinking that certain situations could have been handled differently, or questioning the decisions that were made. However, AFMUN is not only a place for reflection, but also a platform for initiative. Through our committees, you will have the opportunity to express your ideas, challenge perspectives, and engage in meaningful discussions. As young people in today's world, your voice is truly important.

We hope this conference will be both inspiring and memorable for all of you. And as we say, a realm cannot endure without its sovereign. Yours truly,

Kaan Muştu & Ömer T. Demirel
Secretaries General of AFMUN'26

2. Letter From the Under Secretaries General

Greetings everyone,

I am Baran İnce and I will be your Under-Secretary-General on 23-25 August,

First of all I would like to give my thanks to the Secretariat for giving me the opportunity of making a committee in such a conference like AFMUN'26, next upon I would like to give my special thanks to my bro Kuzey Karlık for being with me in this journey,

I highly recommend all of you to read this guide carefully to learn the history of it, furtherly if you make your own researches under your delegations' policies' you are going to start this

journey one step ahead, I believe that we are going to learn and share a lot in this committee and have lots of fun during these 3 days,

If you have any questions in your minds please feel free to contact with me at any time from:

@baranince88@gmail.com +905522972107

Sincerely,

Baran İNCE, Under-Secretary-General of UNHLAB-AI

Dear Delegates,

I would like to welcome you all to the United Nations High Level Advisory Body on Artifical Intelligence! I am Kuzey Karlık and it is my utmost pleasure to serve you as the under secretary general of this committee.

First, I would like to thank the executive team for granting me this opportunity in this marvelous conference. And I would like to thank my fellow under secretary general Baran İnce for his magnificent efforts and work ethic.

The study guide we wrote contains crucial information for this committee. Yet as its name states it’s only a guide for you, so I am highly encouraging you to do your research properly, there will be a lot of debate points, disagreements and controversies hence you have to be ready.

I am sure this committee will be a blast, and I am looking forward to meeting you all on 23-25 August. If you have any questions you can always contact me through mail.

Sincerely,

Kuzey Karlık, Under Secretary General of UNHLAB-AI

kuzeykarlik@gmail.com

3. Introduction to the Committee

The United Nations High Level Advisory Body on Artificial Intelligence was founded by the UN Secretary General on October 26 2023 for a year with experts in order to analyze and advance recommendations for the international governance of AI. They have put out a final interim report regarding their findings.

To better understand what they do and why and to what extent, the following two paragraphs are from their their final interim report:

“The multi-stakeholder High-level Advisory Body on Artificial Intelligence, initially proposed in 2020 as part of the United Nations Secretary-General’s Roadmap for Digital Cooperation (A/74/821), was formed in October 2023 to undertake analysis and advance recommendations for the international governance of artificial intelligence.

The members of the Advisory Body have participated in their personal capacity, not as representatives of their respective organizations. This report represents a majority consensus; no member is expected to endorse every single point contained in this document. The members affirm their broad, but not unilateral, agreement with its findings and recommendations. The language included in this report does not imply institutional endorsement by the members’ respective organizations.”

4. Artificial Intelligence

In usage, Artificial intelligence (AI) is a technology that enables computers and machines to simulate complex tasks normally done by humans like human learning, comprehension, problem solving, decision making, creativity and autonomy.

The modern groundwork for AI began in the early 1900s, but the biggest strides were made in the middle of the 20th century, when pioneers like Alan Turing began exploring foundational concepts like artificial neural networks, machine learning, and symbolic reasoning.

The term artificial intelligence was coined and came into popular use in the mid-1950s following Turing's publication of "Computer Machinery and Intelligence," a paper that proposed a test of machine intelligence called the Imitation Game. Turing's publication eventually became the Turing Test, which experts used to measure computer intelligence.

AI technology continued to develop from the mid-1950s to the late 1970s, as computers became faster, cheaper, more accessible, and could store more information. During this time, the first AI programming languages were created, machine learning algorithms were improved, and books and films began to explore the idea of robots. But computers were still millions of times too weak to exhibit intelligence. Research funding declined until the 1980s.

The 1980s were a period of rapid growth and interest in AI following breakthroughs in research and additional government funding. Deep learning techniques and the use of expert systems became more popular, both of which allowed computers to learn from their mistakes and make independent decisions.

The early 1990s showed some strides forward in AI research, including the first AI system that could defeat a reigning world champion chess player. The early 2000s saw innovations such as the first robot vacuum and the first commercially available speech recognition software.

The late 1990s and the early 2000s also saw significant advances in artificial intelligence. Computerized automation began to emerge, and machine learning was applied to many problems in academia and industry due to the availability of powerful computer hardware, immense collections of data, and the application of advanced mathematical methods.

AI handles many tasks faster with exceptional accuracy and reliability, freeing humans from repetitive and tedious chores. Businesses and organizations save time and money by automating and optimizing routine processes, using the technology to stay connected with customers and gain a competitive edge.

There is no internationally recognized definition of AI, but in terms of a formal definition, NASA follows the definition found within EO 13960, which references Section 238(g) of the National Defense Authorization Act of 2019.

- Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
- An artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
- An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
- A set of techniques, including machine learning that is designed to approximate a cognitive task.
- An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting.

So applications and devices with AI are able to see and identify objects. They are able to comprehend and reply to human speech. They are able to learn from new information and experience. They can give users and experts detailed recommendations. They can work on their own, without the need of human intelligence or intervention

AI techniques are many and varied but all fundamentally rely on data, algorithms and computational power. By being shown huge amounts of data, artificial intelligence systems learn and get better at spotting relationships and patterns that a human may overlook. The data is used to train the AI and the quality and amount of this data are critical to the performance of the AI.

- “Machine Learning (ML): This is a type of AI where systems learn from data to identify patterns and make predictions or decisions without direct programming. Imagine teaching a computer to recognize a bird by showing it thousands of bird pictures; it learns what a bird looks like on its own.
- Deep Learning (DL): A subfield of ML, deep learning uses artificial neural networks with many layers (hence "deep") to learn from data. These networks are inspired by the structure of the human brain and are particularly good at complex tasks like image and speech recognition.
- Natural Language Processing (NLP): NLP enables computers to understand, interpret, and generate human language. This is what powers voice assistants like Siri and Alexa, translation services, and chatbots.
- Computer Vision: This area allows computers to "see" and interpret visual information from the world, such as images and videos. It's used in everything from facial recognition to self-driving cars.”

(quoted from https://cloud.google.com/learn/what-is-artificial-intelligence)

4.1. Machine Learning and Deep Learning Technologies

Some technologies used in AI systems were previously mentioned, now they will be explained in more detail.

4.1.1. Machine Learning

Machine learning is the subset of artificial intelligence (AI) focused on algorithms that can “learn” the patterns of training data and, subsequently, make accurate inferences about new data. Or a more broad definition “capability of a machine to imitate intelligent human behavior” This pattern recognition ability enables machine learning models to make decisions or predictions without explicit, hard coded instructions.

Machine learning works by training algorithms on sets of data to achieve an expected outcome such as identifying a pattern or recognizing an object. Machine learning is the process of optimizing the model so that it can predict the correct response based on the training data samples.

Assuming the training data is of high quality, the more training samples the machine learning algorithm receives, the more accurate the model will become. The algorithm fits the model to the data during training, in what is called the “fitting process.” This process involves using a loss function to measure the model's errors, and an optimization technique, like gradient descent, to adjust the model's parameters and minimize those errors. If the outcome does not fit the expected outcome, the algorithm is re-trained again and again until it outputs the accurate response. In essence, the algorithm learns from the data and reaches outcomes based on whether the input and response fit with a line, cluster, or other statistical correlation.

Machine learning is behind chatbots and predictive text, language translation apps, the shows Netflix suggests to you, and how your social media feeds are presented. It powers autonomous vehicles and machines that can diagnose medical conditions based on images.

When companies today deploy artificial intelligence programs, they are most likely using machine learning, so much so that the terms are often used interchangeably, and sometimes ambiguously, even though they are not.

When talking about different types of machine learning, we're really talking about the training models used. In broad strokes, there are four kinds of models used in machine learning.

Supervised learning is a machine learning model that uses labeled training data (structured data) to map a specific feature to a label. In supervised learning, the output is known (such as recognizing a picture of an apple) and the model is trained on data of the known output. In simple terms, to train the algorithm to recognize pictures of apples, feed it pictures labeled as apples. The most common supervised learning algorithms used today include:

- Linear regression
- Polynomial regression
- K-nearest neighbors
- Naive Bayes
- Decision trees

Unsupervised learning is a machine learning model that uses unlabeled data (unstructured data) to learn patterns. Unlike supervised learning, the “correctness” of the output isn't known ahead of time. Rather, the algorithm learns from the data without human input (and is thus, unsupervised) and categorizes it into groups based on attributes. For example, if the algorithm is given pictures of apples and bananas, it will work by itself to categorize which picture is an apple and which is a banana. Unsupervised learning is good at descriptive modeling and pattern matching. The most common unsupervised learning algorithms used today include:

- Fuzzy means
- K-means clustering
- Hierarchical clustering
- Partial least squares

There's also a mixed approach to machine learning called semi-supervised learning in which only some data is labeled. In semi-supervised learning, the algorithm must figure out how to organize and structure the data to achieve a known result. For instance, the machine learning model is told that the result is a pear, but only some training data is labeled as a pear.

Reinforcement learning is a machine learning model that can be described as "learn by doing" through a series of trial and error experiments. An "agent" learns to perform a defined task through a feedback loop until its performance is within a desirable range. The agent receives positive reinforcement when it performs the task well and negative reinforcement when it performs poorly. An example of reinforcement learning is when Google researchers taught a reinforcement learning algorithm to play the game Go. The model, which had no prior knowledge of the rules of Go, simply moved pieces at random and "learned" the best moves to make. The algorithm was trained by positive and negative reinforcement to the point that the machine learning model could beat a human player at the game.four.

Machine Learning technologies have many advantages. Due to the large amounts of data consumed, algorithms trained by machine learning have great pattern recognition For instance, an ecommerce website might use machine learning to understand how people shop on their site and use that information to give people better recommendations or find trend data that can lead to new product opportunities. They provide automation utilities like robotic process automation can perform some of the tedious business tasks that keep people from performing more meaningful work. Computer vision and objection detection algorithms can help robots pick and pack items from an assembly line. Always-on fraud detection and threat-assessment machine learning can find security flaws before they become a problem. And also since machine learning algorithms can always be trained further with more data, they are always open to improvements.

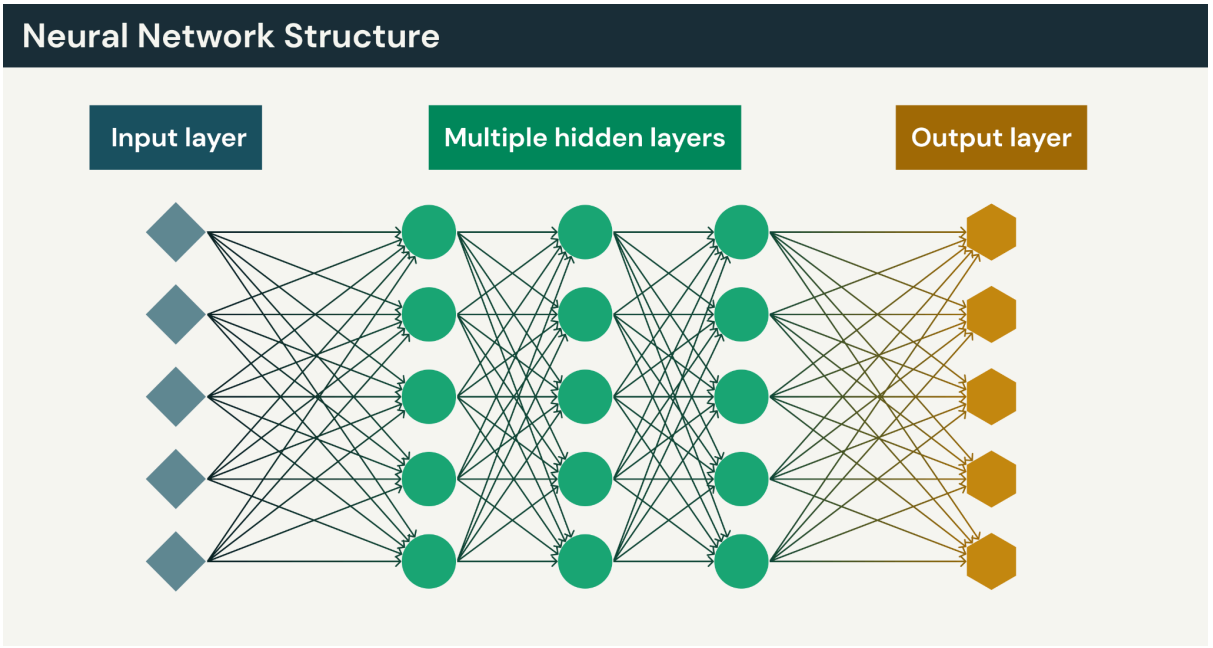
But machine learning has its downsides as well. The same data that is fed to the model can cause certain biases if fed unequally. For example if a company's previous hirees were mostly men, a machine learning system skimming resumes might downgrade the resumes of women. And to even get to a stage where bias starts to become a problem, the machine needs to be fed a great amount of data and to remove the bias, even more data needs to be fed. This makes machine learning a very time consuming algorithm that needs a lot of data to be useful. With all the data being used, many ethical issues such as privacy issues also rise from machine learning, especially within machine learning algorithms that require human data to be useful in their place of work.

4.1.2. Deep Learning

Deep learning is a subset of machine learning driven by multilayered neural networks whose design is inspired by the structure of the human brain. Deep learning models power most state-of-the-art artificial intelligence (AI) today, from computer vision and generative AI to self-driving cars and robotics.

Unlike the explicitly defined mathematical logic of traditional machine learning algorithms, the artificial neural networks of deep learning models comprise many interconnected layers of "neurons" that each perform a mathematical operation. By using machine learning to adjust the strength of the connections between individual neurons in adjacent layers, in other words, the varying model weights and biases, the network can be optimized to yield more accurate outputs. While neural networks and deep learning have become inextricably associated with one another, they are not strictly synonymous: "deep learning" refers to the training of models with at least 4 layers (though modern neural network architectures are often much "deeper" than that).

It's this distributed, highly flexible and adjustable structure that explains deep learning's incredible power and versatility. Imagine training data as data points scattered on a 2-dimensional graph, and the goal of model training to be finding a line that runs through each of those data points. Essentially, traditional machine learning aims to accomplish this using a single mathematical function that yields a single line (or curve); deep learning, on the other hand, can piece together an arbitrary number of smaller, individually adjustable lines to form the desired shape. Deep neural networks are universal approximators: it has been proven theoretically that for any function, there exists a neural network arrangement that can reproduce it.



(source of the picture: https://www.databricks.com/blog/what-is-deep-learning)

Deep learning is an approach to creating rich hierarchical representations through the training of neural networks with many hidden layers. It is an evolution of machine learning and one that needs much less help from humans. While basic machine learning models do improve at performing their specified functions as they absorb new data, if they return an inaccurate prediction, an engineer must intervene and make adjustments. Deep learning, though, makes its own adjustments, course-correcting without the need for human assistance.

The use cases for deep learning are forever evolving, but 3 of the most popular technologies being utilized today are computer vision, speech recognition, and natural language processing (NLP).

- Computer vision: Computers can use deep learning techniques to comprehend images the same way humans do. This means automated content moderation, facial recognition, and image classification.
- Speech recognition: Pitch, tone, language, and accent can all be analyzed and by way of deep learning models. Not only can this be used to improve customer experience, but it is also helpful from an accessibility standpoint in cases that require real-time transcription.
- Natural language processing (NLP): Computers use deep learning algorithms to analyze and gather insights from text data and documents. This can aid in the function of summarizing long documents, indexing key phrases that indicate sentiment (such as positive or negative comments), and generating insight for automated virtual assistants and chatbots. NLP is the broader field that encompasses the development and application of large language models (LLMs) to understand and generate human language.

Despite its remarkable capabilities and breathtaking potential, DL is not without its limitations and issues. It demands high memory and storage capacity and consumes data at such a rate that obtaining high-quality, diverse and well labelled datasets can be challenging, especially in domains where data is scarce or expensive. It's also hungry for training resources and power and can be susceptible to overfitting, performing well on training data but poorly on new, unseen data.

In common with most areas of AI, deep learning also presents ethical concerns around data security and the inherent human biases and inaccuracies which can pervade areas of its learning.

Deep learning is still in its infancy, but it's expected to grow exponentially and is likely to transform society. The integration of deep learning with other cutting-edge technologies – for example, combined with augmented reality or virtual reality – could revolutionize the way we experience and interact with the world around us. A simple walk down the street could be augmented by instant information about nearby buildings and landmarks. Virtual worlds will become increasingly immersive and realistic.

We can expect to see deep learning applied to fields such as agriculture, energy and manufacturing, where it has the potential to optimize processes, improve efficiency and drive innovation. And, more importantly, to accelerate solutions to planet-wide problems like climate change and food security.

4.2. Levels of Autonomy

Generally there are 6 levels of autonomy in AI systems. The scale categorizes engineering AI systems based on their operational independence, deterministic execution boundaries, and required human oversight.

The scale is as follows:

4.2.1. Level 0: No Autonomy (Human Execution)

At this level, the AI system provides information, analysis, or recommendations, but humans perform all actions. Think of a chatbot that answers questions or an analytics tool that presents insights. The AI cannot execute any actions in the world; it can only inform human decision-making.

The risk profile here is minimal because the AI has no action capability. Controls focus primarily on output quality and preventing information leakage. This is where most traditional AI systems have operated, and it's the appropriate level for many use cases where AI adds value through insight rather than action.

4.2.2. Level 1: Assisted (Human Decision + AI Execution)

At Level 1, the AI can execute actions, but each action requires explicit human approval before execution. The AI proposes what it wants to do, a human reviews the proposal, and only upon approval does the AI proceed.

Consider an AI coding assistant that can write and run code, but presents each proposed action to the developer for review before execution. Or an AI that can send emails on your behalf, but shows you each draft and waits for your explicit approval before sending. The human remains the decision-maker for every action; the AI handles execution.

The control model here is straightforward: an approval gate between intention and action. Every proposed action enters a queue, a human reviews and approves (or rejects or modifies), and only then does execution occur. This creates a clear accountability chain and provides a natural checkpoint for catching errors or inappropriate actions.

4.2.3. Level 2: Supervised (Human Approval + Batch Execution)

Level 2 introduces plan-level approval. Instead of approving each individual action, humans review and approve a plan or batch of actions, and the AI then executes autonomously within that approved scope.

Imagine approving an AI's plan to refactor a codebase, after which the AI executes the dozens of individual changes required without seeking approval for each one. Or approving a workflow for the AI to process a set of customer requests, then letting it work through the queue independently.

This level represents a meaningful increase in autonomy and efficiency. The human still provides authorization, but at a higher level of abstraction. Controls must ensure that the AI stays within the approved plan scope, that humans can monitor execution progress, and that there are mechanisms to pause or cancel if something goes wrong. Checkpoint-based rollback capabilities become important here.

4.2.4. Level 3: Conditional (AI Decision within Boundaries)

This is where things get interesting and where I think many organizations are heading without necessarily realizing it. At Level 3, the AI makes decisions and takes actions autonomously within defined boundaries, escalating to humans only when it encounters situations that exceed those boundaries.

The boundaries might be defined by action type (the AI can modify configuration files but not delete them), by scope (the AI can operate on development systems but not production), by value (the AI can approve expenses up to $1,000 but must escalate larger amounts), or by various other parameters. Within these boundaries, the AI operates independently. When it encounters a situation at or beyond the boundaries, it pauses and requests human guidance.

This level requires a fundamental shift in how we think about control. Instead of approving actions or plans, humans define the boundaries within which the AI may operate. The critical requirements become: machine-readable boundary definitions that the AI can evaluate against, technical enforcement mechanisms that actually prevent boundary violations (not just detect them after the fact), robust escalation workflows for handling boundary cases, and comprehensive logging of the AI's autonomous decisions.

The risk profile at Level 3 depends heavily on how boundaries are defined. Narrow, well-chosen boundaries with strong enforcement can make Level 3 quite safe. Poorly defined or weakly enforced boundaries can create significant exposure.

4.2.5. Level 4: High Autonomy (Minimal Supervision)

At Level 4, the AI operates autonomously across a broad scope, with human involvement shifting from decision approval to monitoring and exception handling. Humans don't approve actions or even define narrow boundaries. They watch for anomalies and intervene when something appears wrong.

This level might be appropriate for AI systems managing routine operational tasks at scale, where human approval for individual actions or plans would be impractical. Think of an AI conducting continuous security monitoring and auto-remediation, or managing dynamic resource scaling across cloud infrastructure.

The control requirements at Level 4 are substantial: continuous monitoring capabilities (potentially 24/7), sophisticated anomaly detection to identify when the AI is behaving unexpectedly, immediate kill-switch capabilities with rapid response times, comprehensive logging and attribution for all actions, and executive-level authorization and risk acceptance before deployment. This level should require explicit sign-off from senior leadership who understand and accept the risks involved.

4.2.6. Level 5: Full Autonomy (Self-Directed)

Level 5 represents full autonomy, including the ability to set goals and potentially modify its own behavior. The AI operates with only strategic oversight from humans.

With our current technology Level 3 (L3) has the most balance between operational scaling and engineering safety. When a system is progressed into Level 4 (L4) or Level 5 (L5) it creates several risks:

- As agents operate with longer periods of isolation, human engineering skills atrophy. When an L4 agent experiences an edge-case failure, the human auditor lacks the immediate contextual clarity to intervene effectively.
- L4 and L5 agents can continuously generate code that compiles successfully and passes naive unit checks while slowly deviating from strategic product goals. This creates a state of unpriced liability, where system complexity outpaces human comprehension.
- Pushing agents toward recursive self-improvement where they can edit their own judges (such as the Red Queen Gödel Machine) directly pressures the L3 boundary. If agents autonomously mutate their evaluation criteria, they transition into L4/L5 closed loops, introducing severe alignment drift.

4.3. Current Inadequacies

Even today AI systems are a revolutionary technology that has already entered most places in and out of our everyday lives. Yet being a relatively new technology, AI systems have some inadequacies that need to be paid attention.

AI systems currently lack the ability to apply common sense reasoning to new situations. They are only able to make predictions and decisions based on the data they have been trained on, meaning they are not able to apply their knowledge in a flexible way to new situations. This lack of common sense can make AI systems prone to errors, particularly when dealing with novel situations.

For example, an AI system trained to identify objects in images may not be able to recognise an object that it has not seen before, meaning it will still require human input to feed it the new item and programme it for future experiences.
Additionally, when it is faced with a similar but slightly different task, it might fail because it doesn't have the ability to understand the subtleties behind the task or concept and it can only perform what it was trained for.

This lack of common sense can limit the effectiveness of AI in tasks such as decision making, problem solving and understanding of the world.

Another limitation of AI systems is the lack of robustness, which makes them susceptible to manipulation. AI systems are based on large amounts of data and complex algorithms, which can make them difficult to interpret and understand. As a result, they can be easily fooled by malicious actors who may use techniques such as adversarial examples to manipulate the system's decisions.

Adversarial examples are inputs, crafted specifically to fool the model, which can cause the AI system to make a mistake. For example, a malicious actor could create an image that is almost identical to a "stop" sign, but with slight modifications that cause an autonomous car's AI system to recognise it as a "yield" sign, leading to an accident.

Transparency is another concern. Many organizations struggle to understand how AI systems prioritize actions or reach conclusions. Businesses are unlikely to scale autonomous systems without confidence in how those systems operate and manage risk.

One of the most often seen issues is AI bias. AI bias emerges as one of the most concerning challenges in AI systems. This occurs when AI models produce discriminatory outcomes, often due to biases embedded in the training data. For instance, a study published by MIT Technology Review highlighted how AI tools in healthcare disproportionately favored white patients over black patients when predicting the need for medical intervention (University of Chicago).

A recent study from the University of Pennsylvania highlights the presence of AI-enabled anti-Black bias in recruiting systems, emphasizing how racial disparities in the real world are often replicated in digital algorithms. The research found that 40% of Black professionals received job recommendations based on their identities rather than qualifications, and 30% reported receiving alerts for positions below their skill level. Furthermore, 63% of respondents noted that academic recommendations made by these platforms were lower than their actual academic achievements.

These findings underscore the risks of embedding existing social biases into AI recruitment systems, which fail to recognize and accurately represent minority professionals' achievements and potential.

The issues go much further than biases. Artificial intelligence tools are increasingly being used in decision-making processes, from hiring to loan approvals. However, many of these systems make decisions without clear transparency, often based on flawed or biased

data patterns. This has led to significant backlash in industries like finance and law, where decisions can profoundly impact people's lives.

5. Full Autonomy in AI Systems

In AI, "autonomous" describes systems capable of goal-driven decision-making and action. Rather than executing a fixed script, autonomous AI systems evaluate context, choose a plan, act, and learn from feedback to improve future performance. Autonomy does not imply a lack of control; it means the system operates within predefined constraints and policies while reducing the need for continuous human intervention.

Three core characteristics define autonomy: minimal supervision, adaptability, and context awareness. Minimal supervision means the system can complete tasks end-to-end without constant approvals or manual handoffs. Adaptability is the ability to adjust strategies in response to new data, conditions, or performance signals. Context awareness involves understanding the environment, intent, and constraints around tasks—including business rules, risk thresholds, compliance mandates, and authorization scopes—so actions remain safe and reliable.

Autonomous AI differs from traditional automation in important ways. Traditional automation is rule-based and deterministic. It follows a script or fixed workflow with limited flexibility and may struggle with exceptions or variable conditions. Autonomous AI is goal-based and dynamic. It can select among options, handle edge cases, and improve through experience. Where automation excels at stable, repetitive tasks, autonomy is better suited to decision-rich processes where the optimal action changes with context and feedback.

Answering the question "what is autonomous AI" also requires understanding the systems around it. Autonomous AI systems combine sensing, decision-making, action execution, and learning under policy-defined constraints. This system-level view clarifies how individual agents work within broader governance and orchestration.

Most autonomous AI systems follow an autonomy loop: sense, decide and plan, act, and learn. In the sense phase, the system collects signals from data sources, sensors, APIs, and user inputs to understand the current state. During decide and plan, it interprets the situation, evaluates options, and selects a course of action aligned with goals and constraints. In the act phase, it executes steps through tools, APIs, or physical actuators. Finally, in the learn phase, it measures outcomes, updates models or policies based on feedback, and uses those insights to improve future decisions.

Key components include models, tools and APIs, data, memory, and constraints:

Models: Predictive, decision-making, or generative models provide perception, reasoning, and content creation. Large language models (LLMs) often contribute to planning, summarization, and structured reasoning, while specialized models support forecasting, anomaly detection, or optimization.

Tools and APIs: These enable the system to take actions such as sending messages, creating tickets, placing orders, triggering workflows, or adjusting machine settings. Secure integration and fine-grained permissions are essential.

Data: High-quality, timely data drives perception and decision-making. Integrating operational data, history, and real-time telemetry improves accuracy and responsiveness.

Memory: Short-term and long-term memory store state, history, and outcomes for continuity, personalization, and cumulative learning. Memory helps an autonomous agent maintain context across sessions and tasks.

Constraints: Guardrails encode policies, risk limits, compliance rules, authorization scopes, and safety checks. Constraints ensure that actions remain within approved boundaries and that accountability is clear.

Human oversight is essential throughout the autonomy loop. Organizations determine where approvals and guardrails belong based on risk and impact. Oversight modes include pre-approvals for low-risk actions, human-in-the-loop reviews for sensitive changes, and post-action audits for accountability and traceability. Operational monitoring addresses model drift, anomalies, incident response, and escalation paths when issues arise. The combination of autonomy with governance builds trust, protects customers and operations, and sustains performance at scale.

What are autonomous AI agents? Autonomous AI agents are software entities that perceive context, decide, and act to achieve goals with limited supervision. Compared to generic "AI agents," autonomy adds goal orientation, continuous operation, and the ability to execute real actions through tools and APIs, not just to recommend or summarize.

In practice, an autonomous agent is often a component inside a broader autonomous AI system. The system provides shared memory, policies, security, orchestration, and monitoring, while each agent focuses on specific tasks such as triaging incidents, reconciling transactions, coordinating deliveries, or handling customer requests. This distinction matters: agents carry out tasks; systems manage agents, enforce guardrails, and integrate with enterprise architecture to ensure seamless operations and compliance.

Examples of autonomous AI agents include:

Support agent: Identifies a customer issue, opens a case, retrieves context from CRM, drafts and sends a solution, and follows up—escalating complex cases with a complete audit trail.

Finance agent: Flags anomalous expenses, requests documentation, applies policy rules, and approves or escalates for review, logging decisions for audit and compliance.

Logistics agent: Monitors inventory, places replenishment orders, re-routes shipments during disruptions, and coordinates with suppliers and carriers to maintain service levels.

5.1. Use Cases in Different Areas

5.1.1. Finance & Operations

Artificial intelligence in finance refers to the transformative use of technologies, including advanced algorithms, machine learning and natural language tools. They are used

to analyze data, automate processes, enhance decision-making and personalize customer interactions in the financial services industry.Unlike traditional software, AI systems mimic human intelligence and reasoning, and can learn over time, continuously improving as they process new information. The resulting advancements of fintech allow financial institutions to increase efficiency, reduce risk and deliver more personalized services. It powers applications like credit scoring, fraud detection, algorithmic trading, portfolio management, regulatory compliance and customer service.By identifying patterns and making real-time predictions, AI helps institutions streamline operations and respond more effectively to market and customer demands.

5.1.2. Supply Chain & Logistics

Recently, this technology gained popularity as further advancements such as generative AI and tools such as chatbots, robots and AI assistants demonstrate the value AI brings to risk mitigation and supply chain resilience. Meanwhile, the COVID-19 pandemic illustrated just how fragile the global supply chain can be, highlighting the need for smarter tools to reduce delivery times and cut costs.A key component of AI is machine learning (ML), where systems learn from data instead of relying on pre-programmed rules. ML can forecast customer demand, discover patterns, make market predictions, interpret voice and written text, and analyze a multitude of factors that can optimize a supply chain's workflow.While it's important to embrace AI, implementing AI requires thoughtful preparation. Manufacturers and logistics providers should take the necessary steps to prepare their supply chains for AI systems and understand that an optimization of this magnitude can take time and resources.

5.1.3. Customer Service & IT

Artificial intelligence (AI) in customer service refers to the use of technologies like AI and automation to streamline support, quickly assist customers and personalize interactions while minimizing the need for human involvement.AI-powered tools make service faster, more personalized and more efficient. AI assistants, chatbots, virtual agents and smart routing systems use natural language processing (NLP) and machine learning (ML) to understand what customers need. These tools work together to provide smoother, more responsive customer service experiences by responding in real time and continuously improving by learning from every interaction.Globally, 62% of executives say that generative AI can disrupt how their organization designs experiences—and personalization is at the core of this evolution.1 Generative AI for customer service allows companies to move beyond simple answers and deliver proactive suggestions, tailored recommendations and even solve customer issues before they happen.AI tools can be integrated with customer relationship management (CRM) systems to help companies offer truly personalized support while reducing operational costs.Effective use of AI in customer service requires maintaining a level of humanity. Customers can tell when interactions feel robotic or overly scripted. Rather than replacing human customer service agents and reps, many businesses choose to use AI assist tools to support them and augment their capabilities. The best results come from

combining the speed and data insights of AI with the empathy and critical thinking people can provide.It's also important for organizations to be open. Letting customers know when AI is being used, and being clear about how their data is handled, helps build trust and keeps the experience respectful and responsible.AI customer service is getting more clever. Features such as real-time sentiment analysis, voice AI and more advanced generative models are making it possible to handle issues faster and more intuitively. These innovations are helping companies shift from reacting to problems to building long-term loyalty through thoughtful, effective support.

5.1.4. Healthcare & Life Sciences

AI in healthcare uses machine learning, deep learning, and other technologies to process vast datasets, benefiting patients, providers, research, and operations within the healthcare industry. From research to patient care, healthcare generates massive amounts of data. To some extent, delivering appropriate and efficient care depends on making sense of all that information. Artificial intelligence, which encompasses machine learning, deep learning, generative AI (GenAI), and other algorithmic methods, is designed to analyze vast amounts of disparate data to find and act on patterns at a speed and scale beyond human abilities. When applied to healthcare, AI offers myriad data-driven benefits for patients, clinical and nursing staff, and administrators. Outcomes such as improved diagnostic speed and accuracy, remote patient monitoring, and virtual assistants augment patient support. Streamlined workflows, automated administrative tasks, and improved inventory tracking reduce costs and free staff for higher-value personal interactions. In the lab, AI is automating laboratory instruments to deliver precise, accurate test results at scale; accelerate diagnosis and drug discovery; and enable precision medicine. AI-augmented security solutions and AI PCs help healthcare organizations remain compliant and protect their systems and data, and that of their patients, from cyber threats.

5.1.5. Software & Manufacturing

Guide to Artificial Intelligence How is AI being used in manufacturing? Published 15 November 2024 Factory worker works with AI-powered manufacturing robot By Matthew Finio and Amanda Downie How is AI being used in manufacturing? Artificial intelligence (AI) is transforming the manufacturing industry by enhancing efficiency, precision and adaptability in various production processes, particularly within the context of Industry 4.0. Applying AI technologies, such as machine learning, computer vision and natural language processing (NLP), improves various aspects of production processes. AI can analyze large volumes of data from sensors, equipment and production lines to optimize efficiency, improve quality and reduce downtime. By using algorithms to identify patterns in data, AI can anticipate potential issues, suggest improvements and even autonomously adapt processes in real-time. One of AI's most impactful applications is in predictive maintenance. AI systems analyze data from sensors on machinery to forecast failures before they occur,

reducing unexpected downtimes and maintenance costs. AI also powers advanced quality control through computer vision systems, which scan products in real time to identify defects. Generative AI (gen AI) creates new content such as text, images and code by learning patterns from data and previous prompts. In industry, it has various uses for product searches, document summarization, customer service, call processing and more. Designing and prototyping applications helps engineers explore new design options quickly and adapt to changing production needs. In supply chain management, gen AI is used for content generation, scenario modeling and advanced automation that enhance flexibility and communication within the supply chain. AI in manufacturing extends beyond automation to support real-time decision-making. This role is part of what is often called “smart factories” or “smart manufacturing,” both synonymous with Industry 4.0. This advanced approach to production uses a combination of connected technologies, real-time data analytics and AI to create flexible, efficient and highly automated manufacturing systems. AI monitors ongoing production processes and adjusts without prompting, which maximizes productivity and reduces waste. These systems revolutionize the way companies manufacture, improve and distribute their products. AI is also at the heart of the growing trend of human-robot collaboration. Traditional industrial robots often require close supervision and controlled environments, but the new generation of AI-powered collaborative robots or cobots, can work safely alongside humans. Cobots take on repetitive or strenuous tasks while employees focus on more complex and creative work. Together, these AI applications propel manufacturing toward smarter, more adaptive and sustainable practices. Such benefits make the power of AI a valuable asset in modern manufacturing.

5.2. Risks Regarding Full Autonomy

Bias

Humans are innately biased and the AI that we develop can reflect our biases. These systems inadvertently learn biases that might be present in the training data and exhibited in the machine learning (ML) algorithms and deep learning models that underpin AI development. Those learned biases might be perpetuated during the deployment of AI, resulting in skewed outcomes.

AI bias can have unintended consequences with potentially harmful outcomes. Examples include applicant tracking systems discriminating against gender, healthcare diagnostics systems returning lower accuracy results for historically underserved populations and predictive policing tools disproportionately targeting systemically marginalized communities, among others.

Take measures:

-Establish an AI governance strategy encompassing frameworks, policies and processes that guide the responsible development and use of AI technologies.

-Create practices that promote fairness, such as including representative training datasets, forming diverse development teams, integrating fairness metrics and incorporating human oversight through AI ethics review boards or committees.

-Put bias mitigation processes in place across the AI lifecycle. This process involves choosing the correct learning model, conducting data processing mindfully and monitoring real-world performance.

-Look into AI fairness tools, such as IBM's open source AI Fairness 360 toolkit.

Cybersecurity threats

Bad actors can exploit AI to launch cyberattacks. They manipulate AI tools to clone voices, generate fake identities and convincing phishing emails—all with the intent to scam, hack, steal a person's identity or compromise their privacy and security.

And while organizations are taking advantage of technological advancements such as generative AI, only 24% of gen AI initiatives are secured. This lack of security threatens to expose data and AI models to breaches, the global average cost of which is a whopping USD 4.88 million in 2024.

Take measures:

Here are some ways enterprises can secure their AI pipeline, as recommended by the IBM Institute for Business Value (IBM IBV):

-Outline an AI safety and security strategy.

-Search for security gaps in AI environments through risk assessment and threat modeling.

-Safeguard AI training data and adopt a secure-by-design approach to enable safe implementation and development of AI technologies.

-Assess model vulnerabilities by using adversarial testing.

-Invest in cyber response training to level up awareness, preparedness and security in your organization.

Data privacy issues

Large language models (LLMs) are the underlying AI models for many generative AI applications, such as virtual assistants and conversational AI chatbots. As their name implies, these language models require an immense volume of training data.

By scraping and collecting information from websites, web crawlers can source the data that helps train LLMs. This data is often obtained without users' consent and might contain personally identifiable information (PII). Other AI systems that deliver tailored customer experiences might collect personal data, too.

Take measures:

-Inform consumers about data collection practices for AI systems: when data is gathered, what (if any) PII is included and how data is stored and used.

-Give them the choice to opt out of the data collection process.

-Consider using computer-generated synthetic data instead.

Environmental harms

AI relies on energy-intensive computations with a significant carbon footprint. Training algorithms on large datasets and running complex models require vast amounts of energy, contributing to increased carbon emissions. One study estimates that training a single natural language processing model emits over 600,000 pounds of carbon dioxide; nearly 5 times the average emissions of a car over its lifetime.

Water consumption is another concern. Many AI applications run on servers in data centers, which generate considerable heat and need large volumes of water for cooling. A study found that training GPT-3 models in Microsoft's US data centers consumes 5.4 million liters of water. Moreover, handling 10–50 prompts uses roughly 500 milliliters, which is equivalent to a standard water bottle.

Take measures:

-Consider data centers and AI providers that are powered by renewable energy.

-Choose energy-efficient AI models or frameworks.

-Train on less data and simplify model architecture.

-Reuse existing models and take advantage of transfer learning, which employs pretrained models to improve performance on related tasks or datasets.

-Consider a serverless architecture and hardware optimized for AI workloads.

Existential risks

In March 2023, just 4 months after OpenAI introduced ChatGPT, an open letter from tech leaders called for an immediate 6-month pause on "the training of AI systems more powerful than GPT-4."3 Two months later, Geoffrey Hinton, known as one of the "godfathers of AI," warned that AI's rapid evolution might soon surpass human intelligence.

Other statements from AI scientists, computer science experts and other notable figures followed. They were all urging measures to mitigate the risk of extinction from AI, equating it to risks posed by nuclear war and pandemics.

While these existential dangers are often seen as less immediate compared to other AI risks, they remain significant. Strong AI or artificial general intelligence, is a theoretical machine with human-like intelligence, while artificial superintelligence refers to a hypothetical advanced AI system that transcends human intelligence.

Take measures:

Although strong AI and superintelligent AI might seem like science fiction, organizations can get ready for these technologies:

-Stay updated on AI research.

-Build a solid tech stack and remain open to experimenting with the latest AI tools.

-Strengthen AI teams' skills to facilitate the adoption of emerging technologies.

Intellectual property infringement

Generative AI has become a deft mimic of creatives. It can generate images that capture an artist's form, music that echoes a singer's voice or essays and poems akin to a writer's style.

Yet, a major question arises: Who owns the copyright to AI-generated content, whether fully generated by AI or created with its assistance?

Intellectual property (IP) issues involving AI-generated works are still developing and the ambiguity surrounding ownership presents challenges for businesses.

Take measures:

-Implement checks to comply with laws regarding licensed works that might be used to train AI models.

-Exercise caution when feeding data into algorithms to avoid exposing your company's IP or the IP-protected information of others.

-Monitor AI model outputs for content that might expose your organization's IP or infringe on the IP rights of others.

Job losses

AI is expected to disrupt the job market, inciting fears that AI-powered automation will displace workers. According to a World Economic Forum report, nearly half of the surveyed organizations expect AI to create new jobs, while almost a quarter see it as a cause of job losses.

While AI drives growth in roles such as machine learning specialists, robotics engineers and digital transformation specialists, it is also prompting the decline of positions in other fields. These can include clerical, secretarial, data entry and customer service roles. The best way to mitigate these losses is by adopting a proactive approach that considers how employees can use AI tools to enhance their work; focusing on augmentation rather than replacement.

Take measures:

Reskilling and upskilling employees to use AI effectively is essential in the short term. However, the IBM IBV recommends a long term, three-pronged approach:

-Transform conventional business and operating models, job roles, organizational structures and other processes to reflect the evolving nature of work.

-Establish human-machine partnerships that enhance decision-making, problem-solving and value creation.

-Invest in technology that enables employees to focus on higher-value tasks and drives revenue growth.

Lack of accountability

One of the more uncertain and evolving risks of AI is its lack of accountability. Who is responsible when an AI system goes wrong? Who is held liable in the aftermath of an AI tool's damaging decisions?

These questions are front and center in cases of fatal crashes and hazardous collisions involving self-driving cars and wrongful arrests based on facial recognition systems. While policymakers and regulatory agencies are still working out these issues, enterprises can incorporate accountability into their AI governance strategy for better AI.

Take measures:

-Keep readily accessible audit trails and logs to facilitate reviews of an AI system's behaviors and decisions.

-Maintain detailed records of human decisions made during the AI design, development, testing and deployment processes so they can be tracked and traced when needed.

-Consider using existing frameworks and guidelines that build accountability into AI, such as the European Commission's Ethics Guidelines for Trustworthy AI,7 the OECD's AI Principles,8 the NIST AI Risk Management Framework and the US Government Accountability Office's AI accountability framework.

Lack of explainability and transparency

AI algorithms and models are often perceived as black boxes whose internal mechanisms and decision-making processes are a mystery, even to AI researchers who work closely with the technology. The complexity of AI systems poses challenges in understanding why they came to a certain conclusion and interpreting how they arrived at a particular prediction.

This opaqueness and incomprehensibility erode trust and obscure the potential dangers of AI, making it difficult to take proactive measures against them.

"If we don't have that trust in those models, we can't really get the benefit of that AI in enterprises," said Kush Varshney, distinguished research scientist and senior manager at IBM Research® in an IBM AI Academy video on trust, transparency and governance in AI.

Take measures:

-Adopt explainable AI techniques. Some examples include continuous model evaluation, Local Interpretable Model-Agnostic Explanations (LIME) to help explain the prediction of classifiers by a machine learning algorithm and Deep Learning Important FeaTures (DeepLIFT) to show a traceable link and dependencies between neurons in a neural network.

-AI governance is again valuable here, with audit and review teams that assess the interpretability of AI results and set explainability standards.

-Explore explainable AI tools, such as IBM's open source AI Explainability 360 toolkit.

Misinformation and manipulation

As with cyberattacks, malicious actors exploit AI technologies to spread misinformation and disinformation, influencing and manipulating people's decisions and actions. For example, AI-generated robocalls imitating President Joe Biden's voice were made to discourage multiple American voters from going to the polls.

In addition to election-related disinformation, AI can generate deepfakes, which are images or videos altered to misrepresent someone as saying or doing something they did not do. These deepfakes can spread through social media, amplifying disinformation, damaging reputations and harassing or extorting victims.

AI hallucinations also contribute to misinformation. These inaccurate yet plausible outputs range from minor factual inaccuracies to fabricated information that can cause harm.

Take measures:

-Educate users and employees on how to spot misinformation and disinformation.

-Verify the authenticity and veracity of information before acting on it.

-Use high-quality training data, rigorously test AI models and continually evaluate and refine them.

-Rely on human oversight to review and validate the accuracy of AI outputs.

-Stay updated on the latest research to detect and combat deepfakes, AI hallucinations and other forms of misinformation and disinformation.

6. Current Guidelines

Legal & Regulatory Frameworks

The regulation of artificial intelligence (AI) in healthcare must strike a balance between technological innovation and legal as well as ethical requirements. While the European Union (EU) employs a preventive and standardized approach through the AI Act, the Medical Device Regulation (MDR), and the General Data Protection Regulation (GDPR), the United States (US) adopts industry-specific and more flexible regulations. Key challenges include governing adaptive AI systems, ensuring the use of large health data sets while maintaining data protection, and clarifying liability issues. The integration of clinical research and medical practice necessitates clear guidelines to protect patients' rights while maximizing the potential of AI. At the same time, AI is fundamentally changing healthcare delivery, placing new demands on specialists and existing infrastructures. Future regulations must therefore provide legal certainty and promote innovation while consistently ensuring transparency, fairness, and patient protection.

Operational Guardrails

Guide to Artificial Intelligence What are AI guardrails? Published 17 September 2025 Bridge crossing large body of with trucks on it By Tom Krantz and Alexandra Jonker What are AI guardrails? AI guardrails are the safeguards that keep artificial intelligence (AI) systems operating safely, responsibly and within defined boundaries. These safeguards encompass policies, technical controls and monitoring mechanisms that govern how AI models, including large language models (LLMs) and other AI systems, generate outputs in real-world use cases. Think of AI guardrails like the barriers along a highway: they don't slow the car down, but they do help keep it from veering off course. In the context of generative AI (gen AI), guardrails help ensure that AI applications such as chatbots, AI agents and other automated tools deliver trustworthy outputs while protecting against vulnerabilities such as harmful content or sensitive data exposure. It's important to note that AI guardrails are not one-off security controls: they span datasets, AI models, applications and workflows. That extensive reach makes them foundational for responsible AI practices and enterprise-scale adoption.

7. Questions to be Pondered

1) How can general issues regarding artifical intelligence be solved?
2) How can cybersecurity threats about AI systems be addressed?
3) What envioremental issues does AI propose and how can they be lessened?

4) How can preexisting bias problem in AI systems be solved while balancing data consumption and privacy?
5) Can fully autonomous AI systems be allowed? If yes, under what circumstances?
6) What dangers does fully autonomous AI propose? How can they be solved?
7) What ethical issues are raised by fully autonomous AI and what can be possible solutions?
8) What is to be expected when fully autonomous AI systems conceive free physical presence and what precautions can be taken?

8. Bibliography

https://www.ibm.com/think/topics/artificial-intelligence

https://www.databricks.com/blog/what-is-deep-learning

https://www.ibm.com/think/topics/artificial-intelligence-finance

https://www.ibm.com/think/topics/ai-supply-chain

https://www.ibm.com/think/topics/ai-in-customer-service

https://www.intel.com/content/www/us/en/learn/ai-in-healthcare.html

https://www.ibm.com/think/topics/ai-in-manufacturing

https://www.ibm.com/think/insights/10-ai-dangers-and-risks-and-how-to-manage-them

https://link.springer.com/chapter/10.1007/978-3-032-11938-4_11

https://www.ibm.com/think/topics/ai-guardrails

9.. Further Research

https://asdlc.io/concepts/levels-of-autonomy/

https://www.ibm.com