Machine Learning and Deep Learning
Lecture-15

By: Somnath Mazumdar
Assistant Professor
sma.digi@cbs.dk

Overview

- Responsible Artificial Intelligence
- Ethical issues in AI
- AI alignment
- NN and It's issues (Tips)

Concerns about trust-related issues specifically as barriers to investment. Their top concerns are cybersecurity (57%), privacy (51%) and accuracy (47%) -- IBM Institute for Business Value generative AI survey

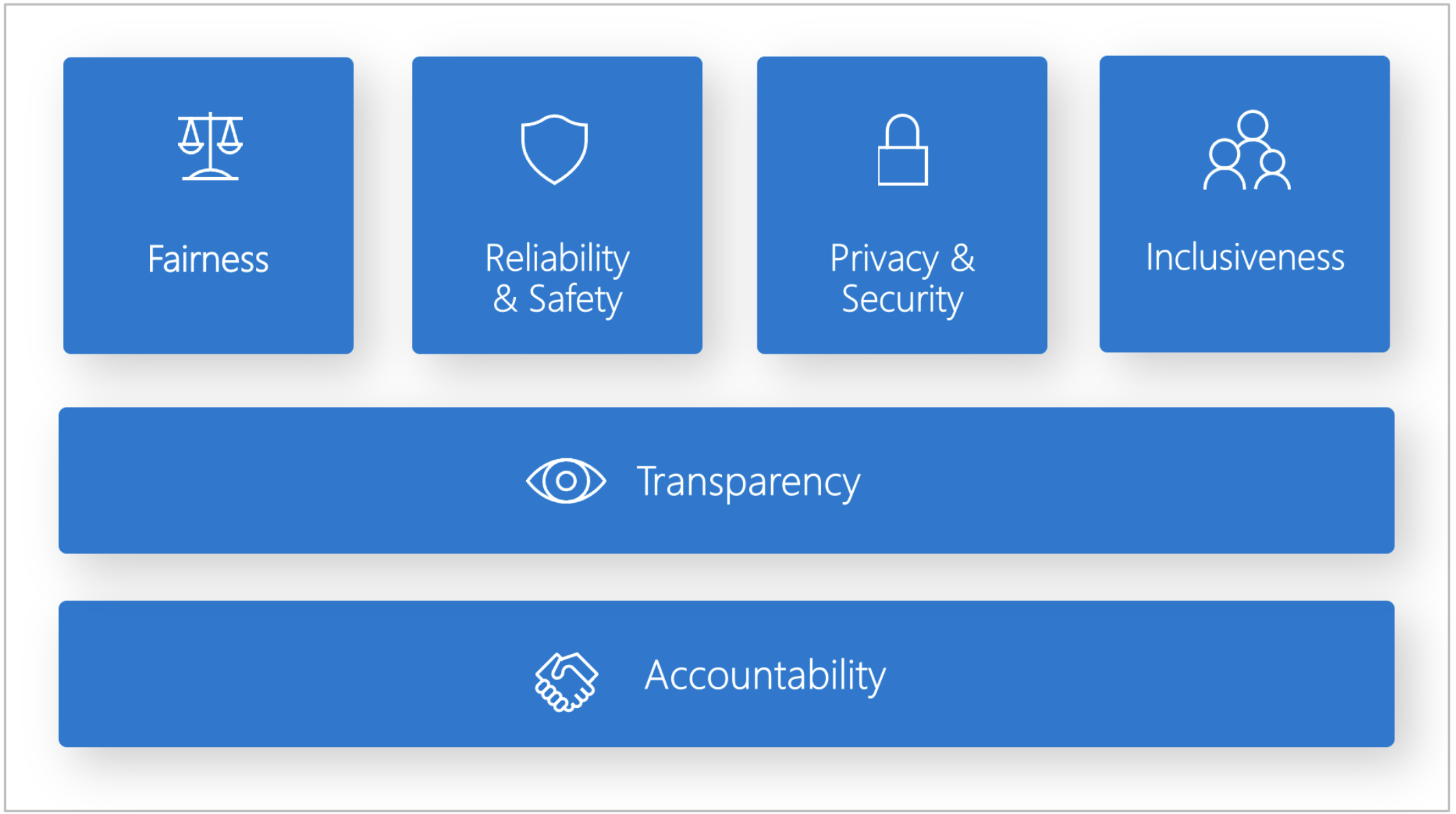
Responsible artificial intelligence (RAI)

RAI

- A set of principles to guide design, development, deployment and use of AI building trust in AI solutions.
- Aim: To embed ethical principles into AI applications.
- Pillars of Trust: Explainability and interpretability are essential for developing trustworthy AI.
  - Explainability: Prediction accuracy + Traceability + Decision understanding.
  - Fairness:
    - Diverse and representative data.
    - Bias-less algorithms/models.
    - Ethical AI review board establishment.
  - Robustness: Handles exceptional conditions (input abnormalities, malicious attacks).
  - Transparency: How service works, evaluate its functionality, and comprehend its strengths and limitations.
  - Privacy: Regulatory frameworks (GDPR).

https://www.ibm.com/topics/responsible-ai

Six Principles of RAI



https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai?view=azureml-api-2

RAI

- Implementing responsible AI practices:
  - Define responsible AI principles
  - Educate and raise awareness in employees
  - Integrate ethics across the AI development lifecycle
  - Protect user privacy

THE HINDU businessline.

FREE TRIAL

Companies / Markets / Portfolio / Opinion / Elections 2024

Home » Info-tech

ChatGPT bug. OpenAI admits data breach at ChatGPT, private data of premium users exposed

Updated - March 25, 2023 at 09:29 AM. | Hyderabad, March 25

According to OpenAI, chat history, names, email addresses, payment addresses, the last four digits of credit cards were exposed

BY K V KURMANATH

COMMENTS

READ LATER

Forbes

BREAKING

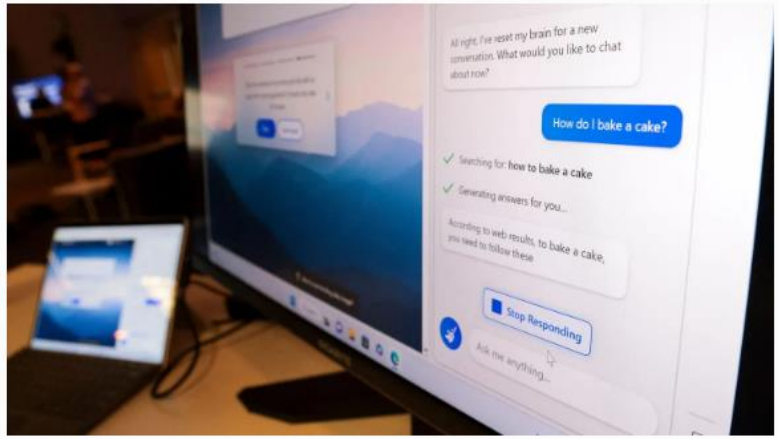
Bing Chatbot's 'Unhinged' Responses Going Viral

Siladitya Ray Forbes Staff
Covering breaking news and tech policy stories at Forbes.

Follow

Feb 16, 2023, 09:15am EST

TOPLINE Microsoft Bing's chatbot has reportedly been sending out strange responses to certain user queries that include factual errors, snide remarks, angry retorts and even bizarre comments about its own identity, just weeks after the company launched to much fanfare an updated version of the search engine which implements OpenAI's ChatGPT technology.



Microsoft Bing search engine in pictured on a monitor in the Bing Experience Lounge during an event ... [+] AFP VIA GETTY IMAGES

KEY FACTS

- Users on the subreddit r/bing have shared examples of the Bing Chatbot's responses to queries that they are calling "unhinged" and "gaslighting" including scenarios where the bot responds angrily to a question or comment and then shares reply prompts that allow the user

Litigation | Attorney Analysis | Data Privacy | Legal Industry

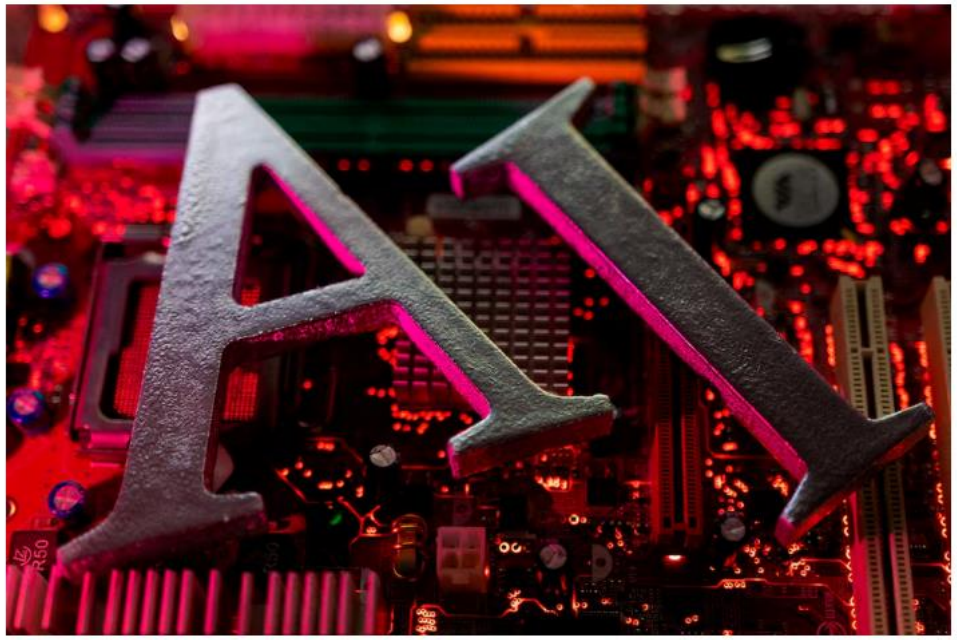
"The perils of dabbling": AI and the practice of law

By **Tony Petruzzi** and **Helena Guye**

September 11, 2023 4:30 PM GMT+2 · Updated 7 months ago

Aa

Commentary | Attorney Analysis from Westlaw Today, a part of Thomson Reuters.



AI (Artificial Intelligence) letters are placed on computer motherboard in this illustration taken June 23, 2023. REUTERS/Dado Ruvic/Illustration/File Photo *Purchase Licensing Rights*

The Guardian

AI and the US election
US elections 2024

This article is more than 8 months old

Disinformation reimagined: how AI could erode democracy in the 2024 US elections

Nick Robins-Early

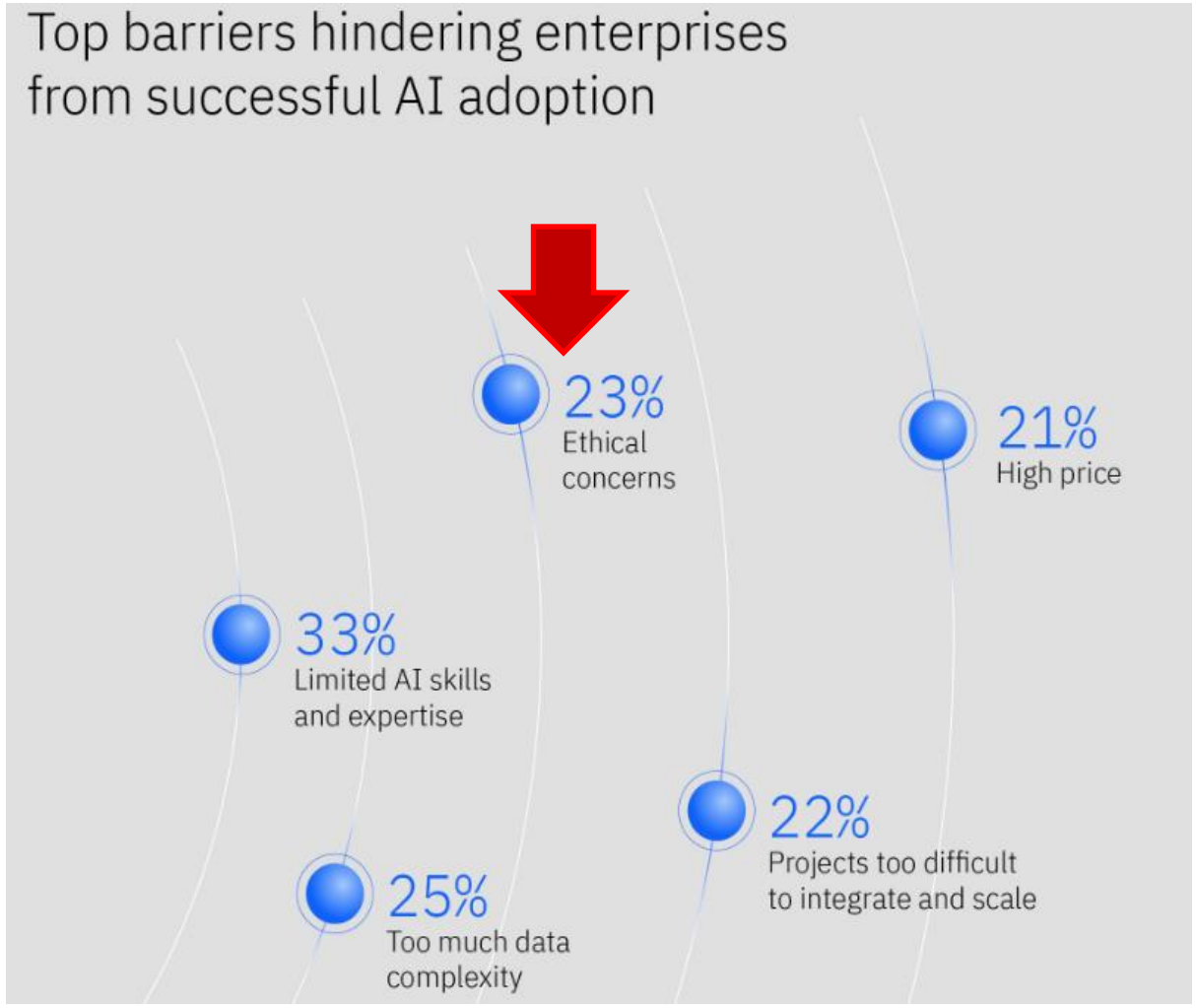
Wed 19 Jul 2023 16.00 CEST

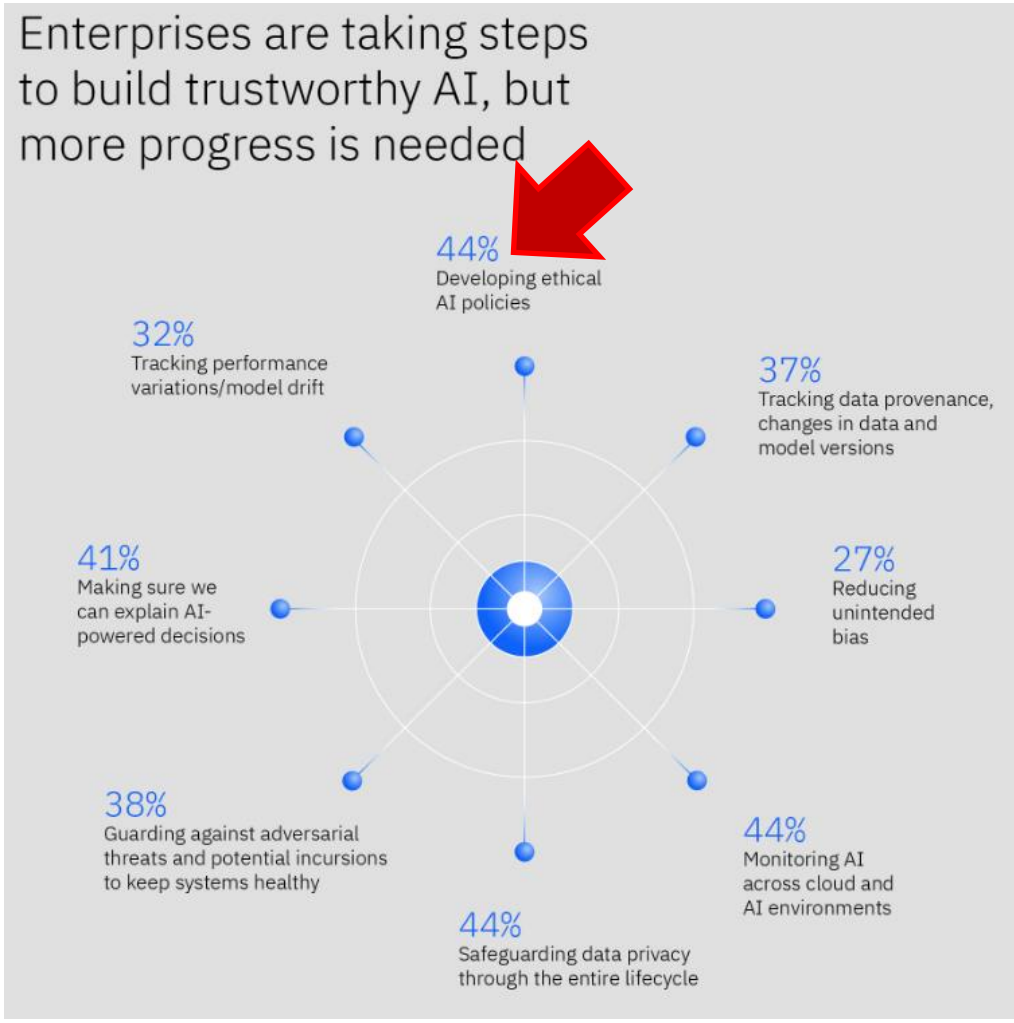
Illustration: Mark Harris/The Guardian

Advances in generative artificial intelligence could supercharge the propaganda playbook, experts warn

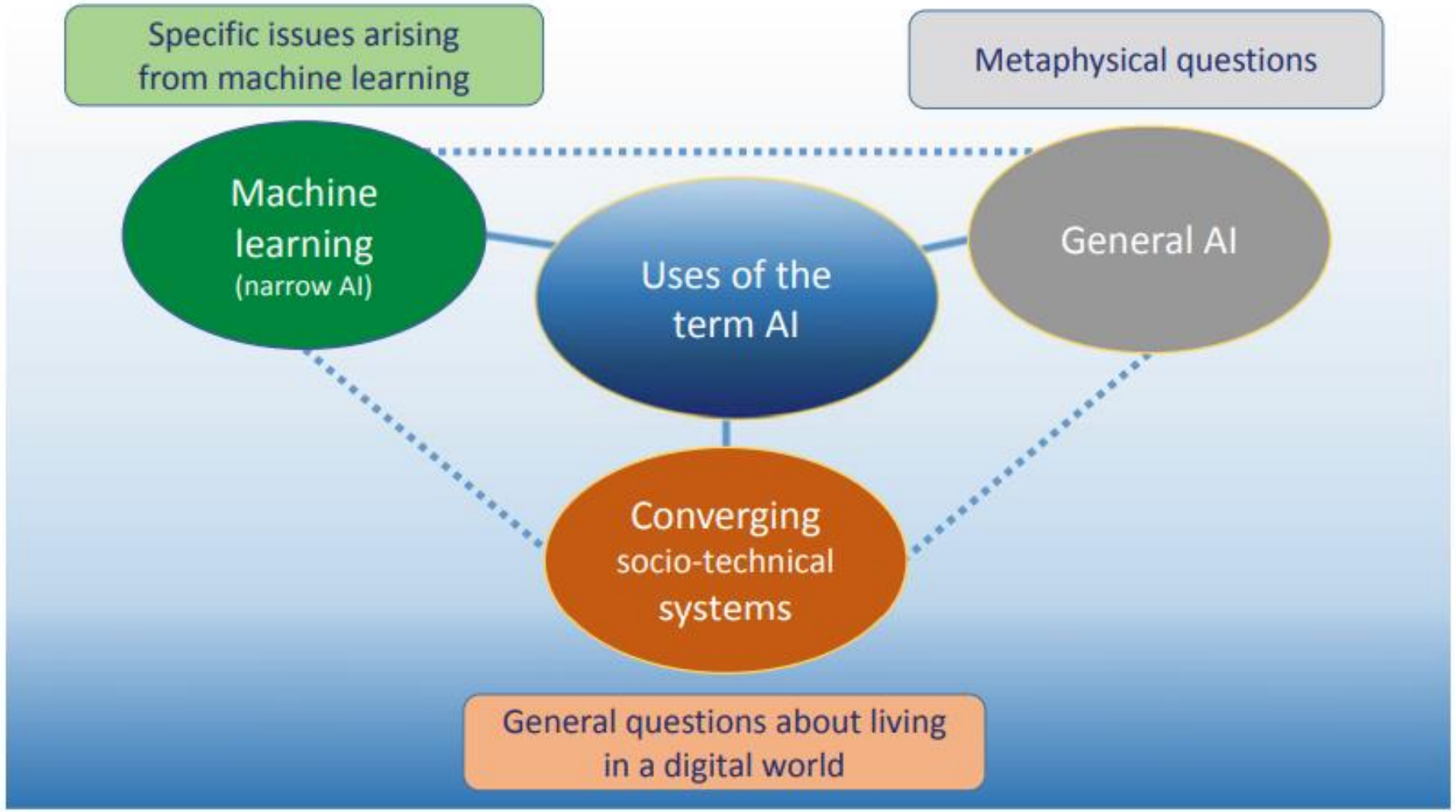
Ethical Issues in AI

IBM Global AI Adoption Index





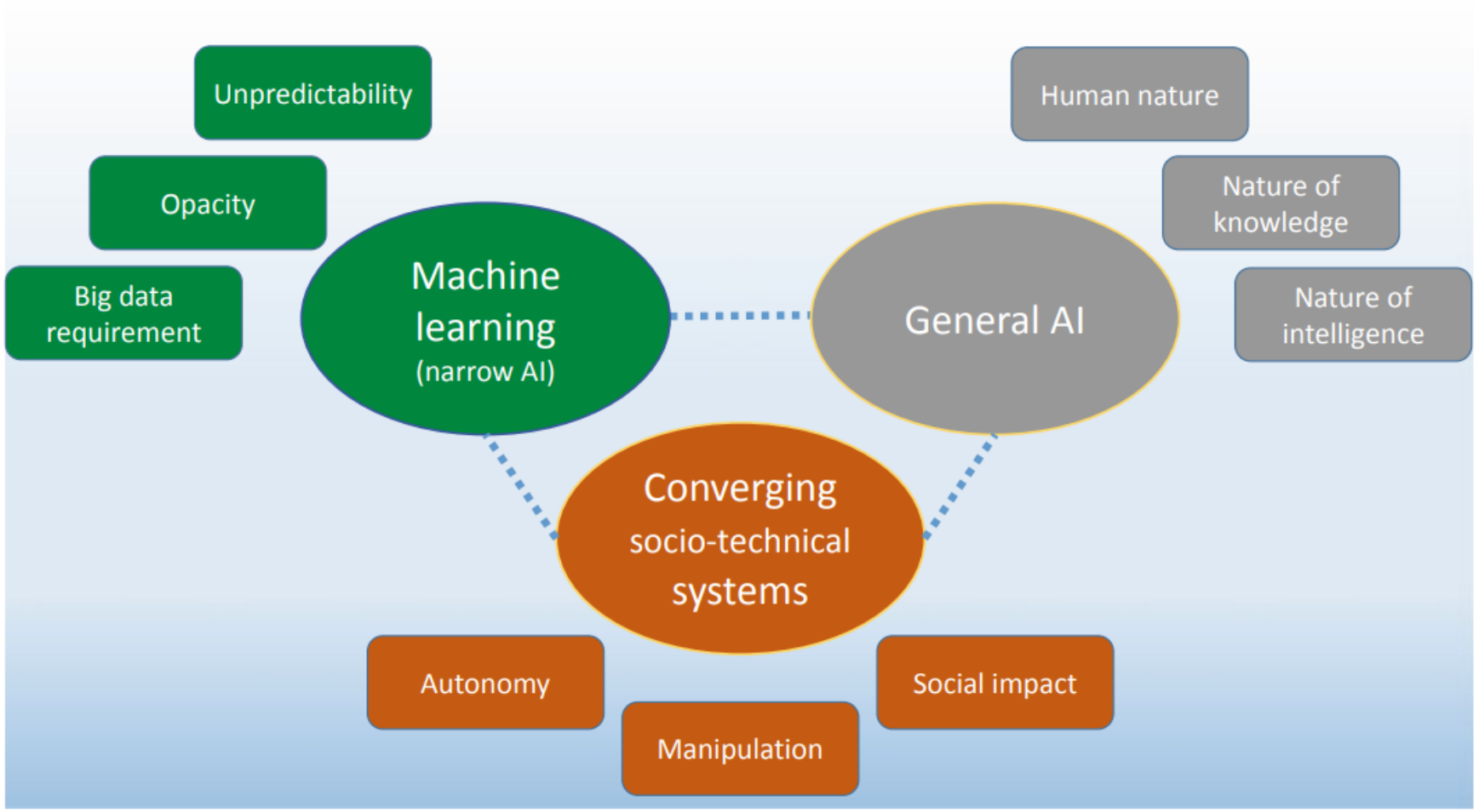
Concepts of AI and Ethical questions they raise



Advanced AI systems

 Evade Shutdown	 Hack Computer Systems	 Make Copies	 Acquire Resources	 Ethics Violation	 Hire or Manipulate Humans	 AI Research & Programming
 Persuasion & Lobbying	 Hide Unwanted Behaviors	 Strategically Appear Aligned	 Escape Containment	 Research & Development	 Manufacturing & Robotics	 Autonomous Weaponry

Key Characteristics of different uses of “AI”



Categories of ethical issues of AI: Issues arising from ML

Privacy and data protection	Lack of privacy
	Misuse of personal data
	Security problems
Reliability	Lack of quality data
	Lack of accuracy of data
	Problems of integrity
Transparency	Lack of accountability and liability
	Lack of transparency
	Bias and discrimination
	Lack of accuracy of predictive recommendations
	Lack of accuracy of non-individual recommendations
Safety	Harm to physical integrity

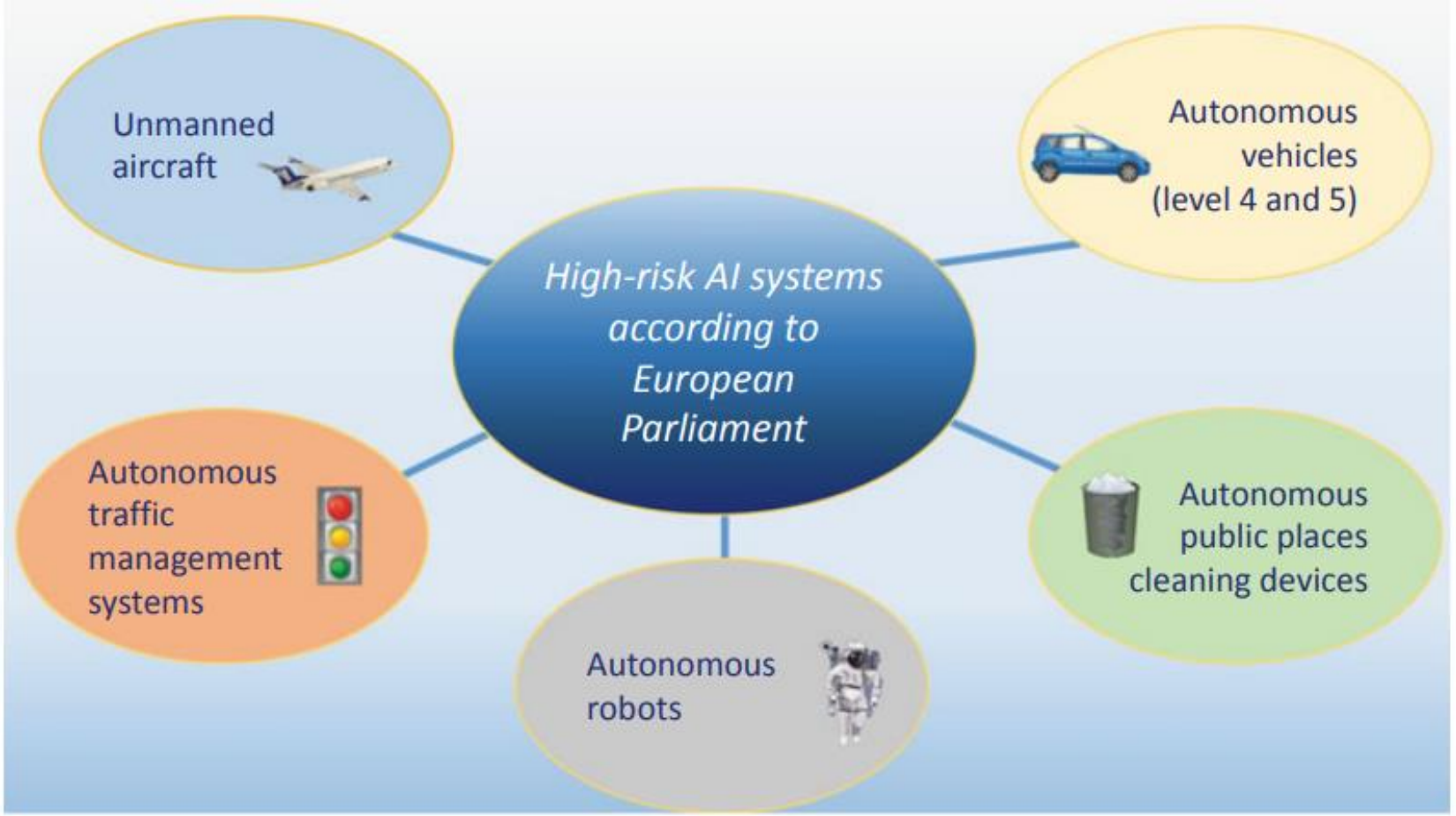
Living in a digital world

Economic issues	Disappearance of jobs
	Concentration of economic power
	Cost to innovation
Justice and fairness	Contested ownership of data
	Negative impact on justice system
	Lack of access to public services
	Violation of fundamental human rights of end users
	Violation of fundamental human rights in supply chain
	Negative impact on vulnerable groups
	Unfairness
Freedom	Lack of access to and freedom of information
	Loss of human decision-making
	Loss of freedom and individual autonomy
Broader societal issues	Unequal power relations
	Power asymmetries
	Negative impact on democracy
	Problems of control and use of data and systems
	Lack of informed consent
	Lack of trust
	Potential for military use
	Negative impact on health
	Reduction of human contact
	Negative impact on environment
Uncertainty issues	Unintended, unforeseeable adverse impacts
	Prioritisation of the "wrong" problems
	Potential for criminal and malicious use

Metaphysical Issues

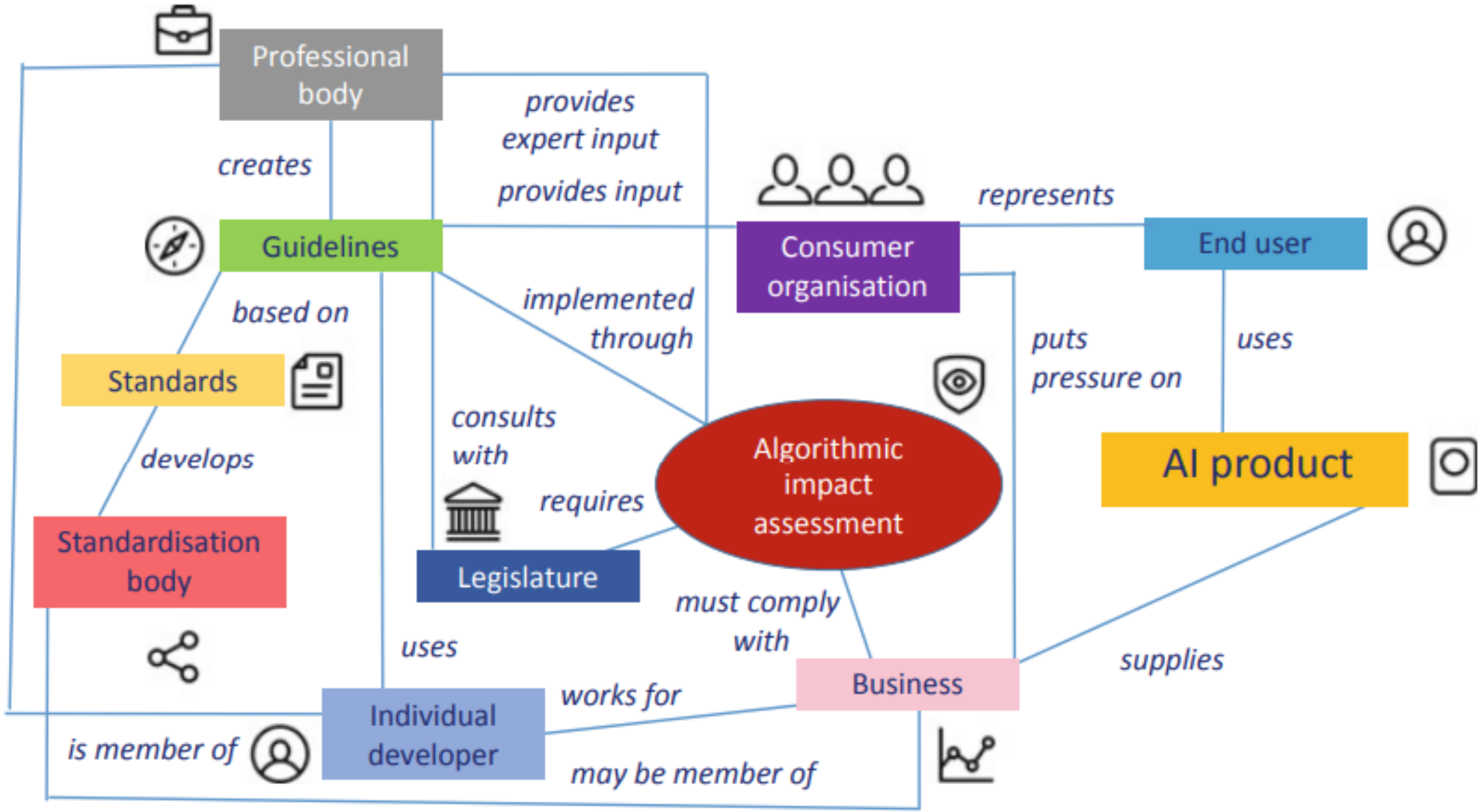
Machine consciousness
“Awakening” of AI
Autonomous moral agents
Super-intelligence
Singularity
Changes to human nature

High-risk AI systems according to European Parliament



Overview of AI Stakeholders

Implementing Algorithmic Impact Assessment

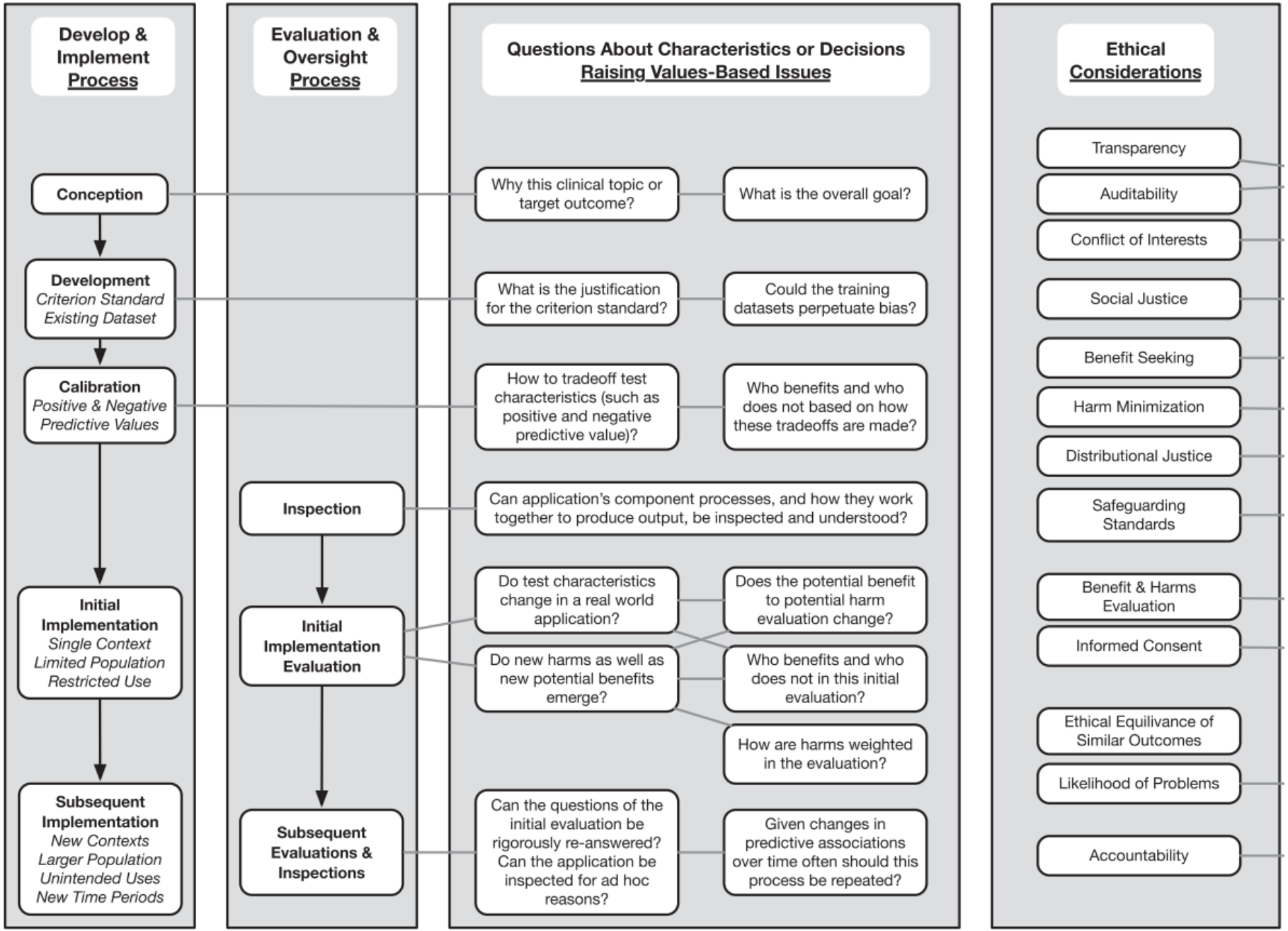


Stahl, B.C. (2021). Ethical Issues of AI. In: Artificial Intelligence for a Better Future. Springer Briefs in Research and Innovation Governance. Springer.

Research integrity			
Component		Examples	Example governance mechanisms
Not engaging in misrepresentation and fraud	2.1	• Fabricating data • Falsifying results • Serious misrepresentation of findings • Conflicts of interest	• Codes of conduct • Peer review • Processes to support/protect whistle blowers • Formal censure and/or future restrictions/bans by publication venues, conferences, professional bodies, institutional home or funding agency • Legal action • Mandatory disclosure of conflicts of interest • Publication venue policies, e.g. not publishing work by actors who stand to benefit from the findings • Pre-registration • Making funding available for research that is independent of interested or possibly conflicting parties
Reproducibility and replicability	2.2	• Reproducibility, e.g. sharing data, code etc. • Replicability, including statistically sound performance claims • Reliability and validity issues • Producing trustworthy results • Proof/evidence of claims • Disclosure of complexity/compute requirements	• Pre-registration • Reproducibility badges • Reproducilibity checklists • Data publication requirements or code Release • Required disclosure of experimental details e.g. compute spending • Peer review, e.g. to scrutinise statistical claims • Funding scientific reproduction projects, including for research that would be prohibitively expensive for non-industry researchers
Addressing assumptions and limitations	2.3	• Disclosure of assumptions, and conditions under which claims hold, including generalisability • Quantifying uncertainty • Discussion of project limitations • Discussion of dataset limitations, including bias	• Standardized reporting practices, e.g. datasheets and model cards • Publication norms and requirements, e.g. mandating a discussion of limitations, error bars on charts, etc. • Peer review, e.g. to scrutinise evidence of claims, and to verify other publication norms and requirements • Research community norms

Use Case: Healthcare Sector

Identifying Ethical Considerations for healthcare



https://doi.org/10.1080/15265161.2020.1819469

AI Alignment

IBM Global AI Adoption Index 2023

- 42% of IT professionals at large organizations actively deployed AI
- 38% of IT professionals at enterprises report they are implementing generative AI

- Automation of IT processes (33%)
- Security and threat detection (26%)
- AI monitoring or governance (25%)
- Business analytics or intelligence (24%)
- Automating processing, understanding, and flow of documents (24%)
- Automating customer or employee self-service answers and actions (23%)
- Automation of business processes (22%)
- Automation of network processes (22%)
- Digital labor (22%)
- Marketing and sales (22%)
- Fraud detection (22%)
- Search and knowledge discovery (21%)
- Human resources and talent acquisition (19%)
- Financial planning and analysis (18%)
- Supply chain intelligence (18%)

What is AI alignment?

- Process of encoding human values and goals into LLMs to make them helpful, safe, and reliable.
- Goal: Enterprises can tailor AI models to follow their business rules and policies.
- Alignment happens during fine-tuning, when a foundation model is fed examples of the target task.
- Involves two steps:
    1. Instruction-tuning phase: LLMs are given examples of target task so it can learn by example.
    2. Critique phase: Human or another AI interacts with the model and grades its responses in real-time.

- E.g., In RL, this step is called RL with human feedback (RLHF) or AI feedback (RLAIF).

https://research.ibm.com/blog/what-is-alignment-ai

Intermediate Objectives: RICE

- Four Key principles: Robustness, Interpretability, Controllability, and Ethicality (RICE).
  1. Robustness: System's stability needs to be guaranteed across various environments.
  2. Interpretability: Operation and decision-making process of the system should be clear and understandable.
  3. Controllability: System should be under guidance and control of humans.
  4. Ethicality: System should adhere to society's norms and values.

Why AI alignment is important?

- Three-step argument: Each step building upon previous one.
  1. DL applications have an increasingly large impact on society and bring significant risks.
  2. Misalignment represents a significant source of risks.
  3. Alignment research and practice address risks stemming from misaligned systems.

https://research.ibm.com/blog/what-is-alignment-ai

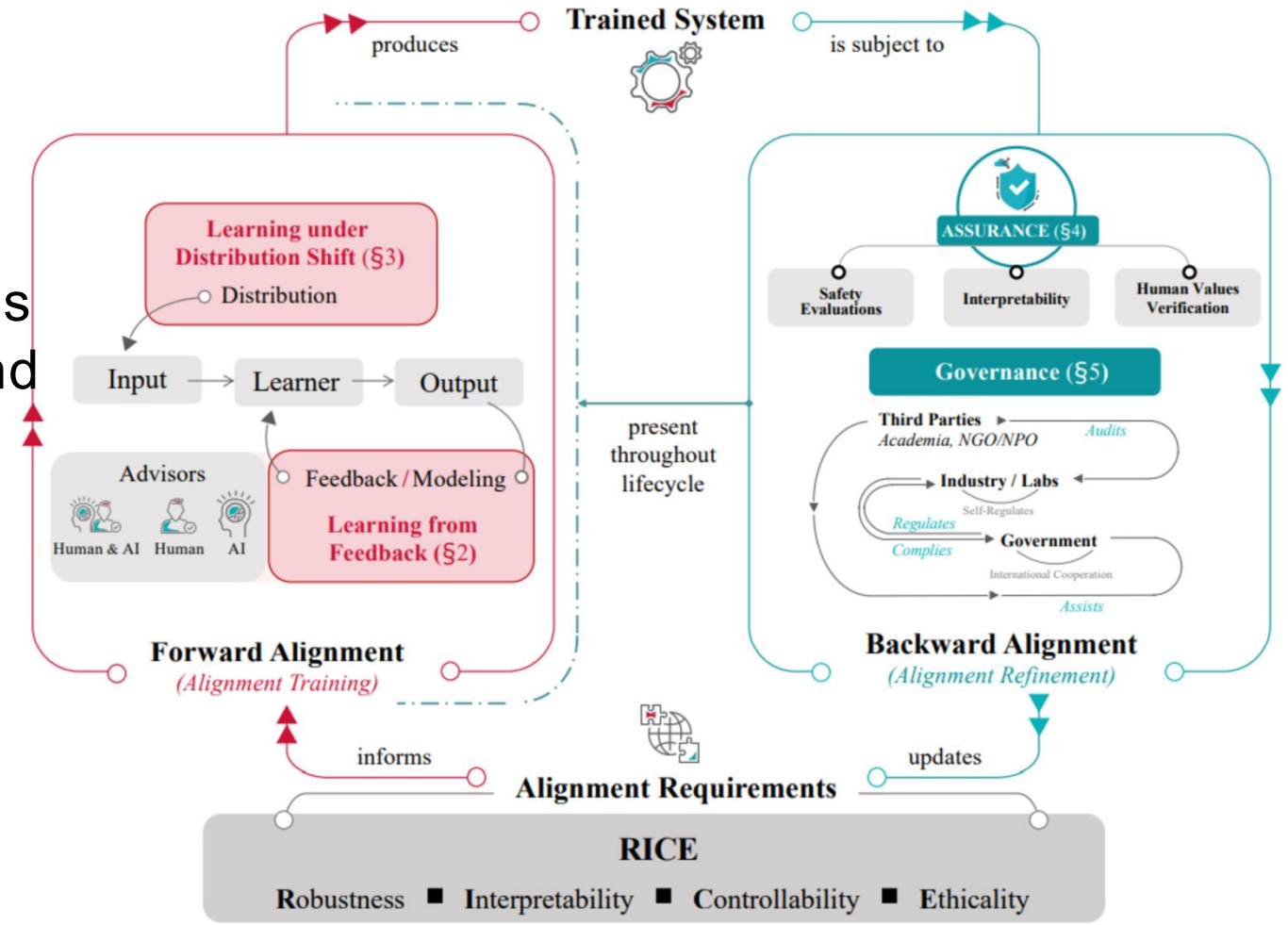
Misalignment

- Two Causes of Misalignment:
  - Reward Hacking: Proxy rewards are easy to optimize and measure, but suffers from capturing full spectrum of actual rewards (misspecified rewards).
    - Optimizing misspecified rewards lead to reward hacking.
  - Goal misgeneralization: Agent actively pursues objectives distinct from training objectives, while retaining capabilities acquired during training.
- Two primary factors of misalignment:
  - Limitations of Human Feedback: During LLMs training
    - inconsistencies can arise from human data annotators.
    - may introduce biases deliberately.
  - Limitations of Reward Modeling:
    - Accurately capturing human values.

https://arxiv.org/pdf/2310.19852.pdf

Alignment Cycle

- Forward Alignment (training) produces trained systems based on alignment requirements.
- Backward Alignment (refinement) ensures practical alignment of trained systems and revises alignment requirements
  - Ensure practical alignment of trained systems.
- Cycle is repeated until reaching a sufficient level of alignment.



https://arxiv.org/pdf/2310.19852.pdf

OpenAI's AI alignment

- OpenAI’s superalignment project is focused on aligning artificial superintelligence systems.
- Focuses on engineering a scalable training signal for smart AI systems that is aligned with human intent. It has three main pillars:
  1. Training AI systems using human feedback.
     - RL from human feedback: InstructGPT[1]
     - Training AI systems to assist human evaluation.
  2. Training AI systems to do alignment research.

1. https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf

NN and It's issues (Tips)

Neural Networks

- If input propagates through the network in only one direction -> feedforward network.
- If there are feedback loops in the network -> Recurrent NN.
- LSTMs retains information for long duration, well suited for time-series prediction.

- First generation of NNs: single-layer NNs and MLPs
- Second generation of NNs: To reduce number of parameters in a network, introduces convolutional NNs (CNNs).
- To solve problem of information loss in CNNs use capsule networks (CapsNets).
- Third generation of NNs: Makes use of spiking NNs (SNNs) to emulate human brain-like functioning.

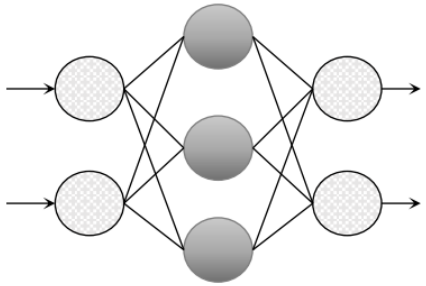
Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." *IEEE Design & Test* 37.2 (2020): 30-57.

Evolution of NN Models

First Generation

Thresholding as Activation

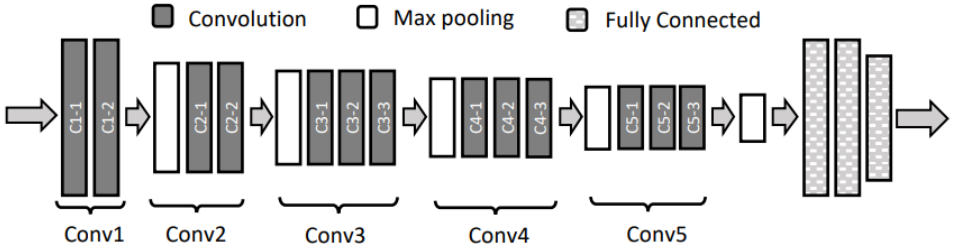
- Single Layer Perceptron
- Multi-Layer Perceptron



Second Generation

Deep Neural Networks

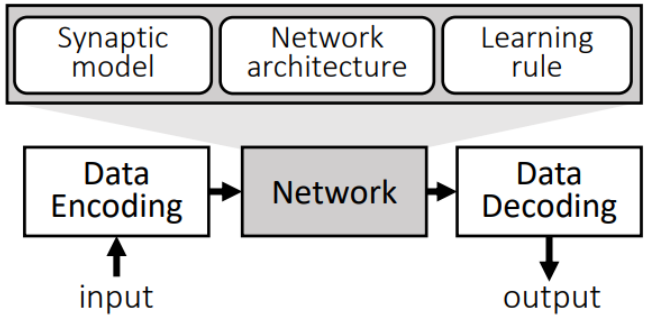
- Convolutional Neural Networks
- Recurrent Neural Networks
- Generative adversarial Network
- Capsule Networks



Third Generation

Spiking Neural Networks

- Deep Spiking Neural Networks



Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." *IEEE Design & Test* 37.2 (2020): 30-57.

Robustness of DNN

- Robustness determines the integrity of the network under varying operating conditions, and the accuracy of DNN outputs in the presence/absence of input or network alterations.
- Divided into two security and reliability.
    1. If attacker cannot steal information, engage system resources, modify network parameters, or render an incorrect input.
    2. If it does not display any changes to its output, parameters, or behavior, due to changes in environmental conditions, during fabrication and deployment.

Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." *IEEE Design & Test* 37.2 (2020): 30-57.

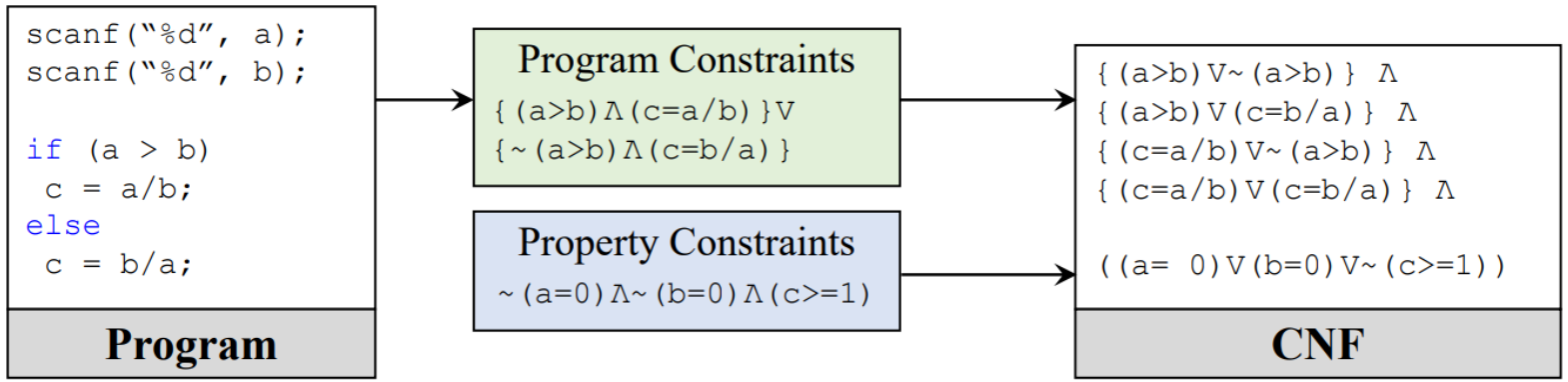
Formal Verification for robust ML

- Verification Categories:
  - Satisfiability (SAT)
  - Linear programming (LP)
  - Satisfiability modulo theories (SMT) solving
  - Theorem proving, and incomplete verification

Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." *IEEE Design & Test* 37.2 (2020): 30-57.

SAT/SMT

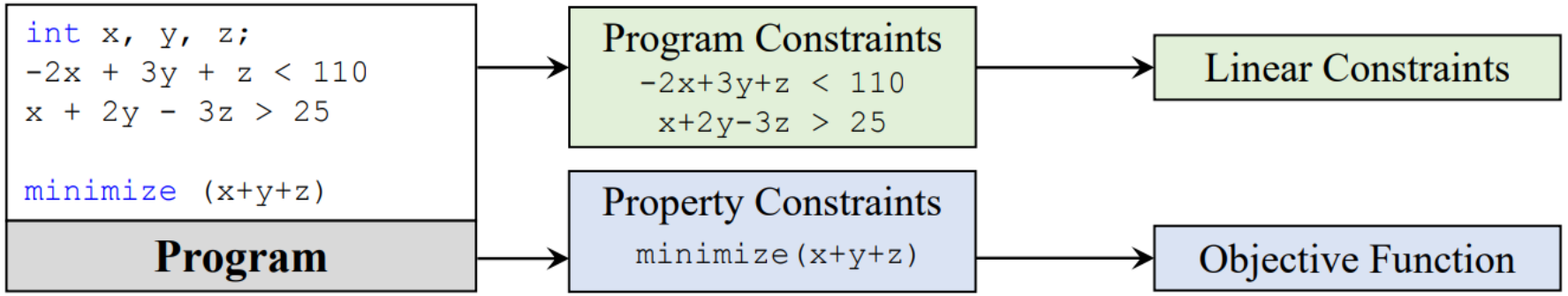
- SAT is a branch of formal verification.
- System model and property to be verified for the system are expressed in a propositional logic, and written into conjunctive normal form (CNF).
- Formula is then checked by an automatic SAT solver.
- SAT-based verification suffers from scalability problem.
- SMT solvers are preferred for verifying DNNs with real and/or integer network parameters.



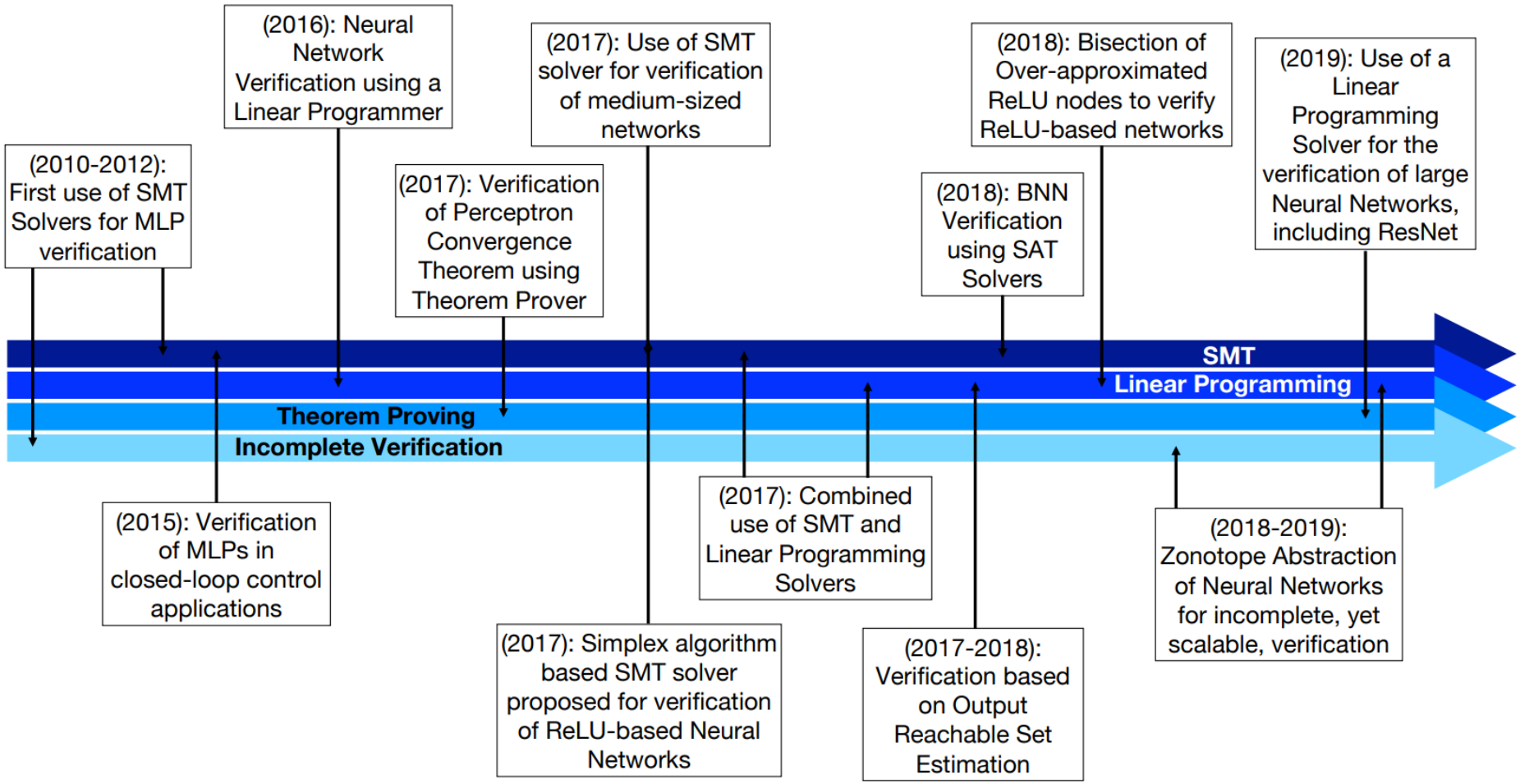
Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." *IEEE Design & Test* 37.2 (2020): 30-57.

Linear programming

- LP-based verification works by defining the system as a set of linear constraints, and property to be verified as an objective function.
- Objective function can be either a minimization or a maximization function.
- DNN verification: LP is used to check robustness of network against adversarial attacks.
- Limitation of LP: Requires linear constraints.



Verification Techniques



Shafique, Muhammad, et al. "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead." *IEEE Design & Test* 37.2 (2020): 30-57.

Issues with ML models

- Class imbalance: Disproportional number of samples per class.
  - Classifier might be biased towards majority classes and achieve bad classification.

- Concept drift: describe problem of changing underlying relationships in data.
  - Approximating a mapping function (f) given input data (x) to predict an output value (y), y = f(x).
    - Digit classification, text categorization or speech recognition, assume that mapping learning from historical data will be valid for new data and relationships between input and output do not change over time. Not true for the problem of malware detection and classification.

Issues with ML models

- Adversarial learning is a technique employed to attempt to fool machine learning by automatically crafting adversarial examples.

- Interpretability: Model is given an input X and it produces an output Y through a sequence of operations hardly understandable to a human.
  - Interpretability of the model determines how easily the analysts can manage and assess the quality and correct the operation of a given model.
    - Example: Cybersecurity analysts prefer rule-based and signature-based systems rather than neural-based methods

Handling ML experiments

- Who created the model at what time?
- Which hyperparameters were used?
- What feature transformations have been applied?
- Which dataset was the model derived from?
- Which dataset was used for computing the evaluation data?

Model Management Challenges

- Model Definition: Difficult to define actual model to manage.
  - Input data needs to be transformed into features expected by model.
  - ML pipelines combines feature transformations and actual model in a single abstraction.

- Model Validation: Ability to back-test accuracy performance of models over time (after model gets update).
  - Make use of same training, test and validation set including same evaluating metrics.

- Model Retraining: Deciding when to retrain a model is challenging.
- Training is done offline and models are loaded at prediction in a system.
- If new events (new public holiday, a new promotional activity) occur that our models have not been trained on, need to trigger retraining or even re-modelling.

Model Management Challenges

- Data Management Challenges: data integration, feature transformation, model training apply different operations based on different abstractions.
- Querying model Metadata to select best model: To accelerate model lifecycle management, need to understand metadata and lineage of models (e.g., hyperparameters, trained and validated datasets).
  - Centralized metadata store can accelerate learning processes (e.g., warm-starting of hyperparameter search) and automate simple error checks.
- Multi-Language Code Bases
- Backwards Compatibility: Model that was trained (last month or last year) should still be working today (required for production deployments).

Tips: Missing Value Handling

- Example: Missing values are problematic for product catalogs of online retailers.
- Manual update is not scalable.
- ML methods are designed for matrices only.
- Data is not available in numeric formats but rather in text/other forms.
  - Sols: Implementing feature extraction steps and imputation algorithms in one single pipeline.
- Must support hyperparameter optimization to automatically perform model selection and parameter tuning.

Recap (1/3)

- Responsible AI (RAI) is a set of principles guiding the design, development, deployment, and use of AI to build trust in AI solutions.
- Aim of RAI is to embed ethical principles into AI applications.
- Pillars of Trust in RAI include explainability and interpretability, which are essential for trustworthy AI, encompassing prediction accuracy, traceability, and decision understanding.
- Fairness in RAI involves diverse and representative data, bias-less algorithms, and the establishment of an ethical AI review board.
- Robustness in RAI means handling exceptional conditions like input abnormalities and malicious attacks.
- Transparency in RAI refers to understanding how a service works, evaluating its functionality, and comprehending its strengths and limitations.
- Privacy in RAI is important, with regulatory frameworks like GDPR being relevant.
- Implementing responsible AI practices includes defining responsible AI principles, educating employees, integrating ethics across the AI development lifecycle, and protecting user privacy.

Recap(2/3)

- Ethical issues in AI are a key concern.
- AI alignment is the process of encoding human values and goals into LLMs to make them helpful, safe, and reliable.
- Goal of AI alignment is to enable enterprises to tailor AI models to follow their business rules and policies. Alignment happens during fine-tuning, involving instruction-tuning and a critique phase.
- Importance of AI alignment stems from the increasing impact and risks of Deep Learning (DL) applications, with misalignment being a significant source of these risks.
- Misalignment has two main causes: reward hacking, where proxy rewards are optimized instead of actual rewards, and goal misgeneralization, where the agent pursues objectives different from training objectives.

Recap(3/3)

- Formal verification methods like SAT, LP, and SMT are used for robust ML.
- Issues with ML models include class imbalance, concept drift, adversarial learning, and a lack of interpretability.
- Handling ML experiments involves tracking model creation details, hyperparameters, feature transformations, and datasets used.
- Model management challenges include defining the model, validation, retraining, and ensuring backwards compatibility.
- Data management challenges involve data integration, feature transformation, and querying model metadata.