Artificial Intelligence

Philosophy and Ethics of AI and the Future of AI

1141AI10
MBA, IM, NTPU (M5276) (Fall 2025)
Tue 2, 3, 4 (9:10-12:00) (B3F17)

https://meet.google.com/paj-zhhj-mya

Min-Yuh Day, Ph.D,
Professor and Director

Institute of Information Management, National Taipei University

https://web.ntpu.edu.tw/~myday

2025-12-09

Syllabus

Week	Date	Subject/Topics
1	2025/09/09	Introduction to Artificial Intelligence
2	2025/09/16	Artificial Intelligence and Intelligent Agents; Problem Solving
3	2025/09/23	Knowledge, Reasoning and Knowledge Representation; Uncertain Knowledge and Reasoning
4	2025/09/30	Case Study on Artificial Intelligence I
5	2025/10/07	Machine Learning: Supervised and Unsupervised Learning; The Theory of Learning and Ensemble Learning

Syllabus

Week	Date	Subject/Topics
6	2025/10/14	NVIDIA Fundamentals of Deep Learning I: Deep Learning; Neural Networks
7	2025/10/21	NVIDIA Fundamentals of Deep Learning II: Convolutional Neural Networks; Data Augmentation and Deployment
8	2025/10/28	Self-Learning
9	2025/11/04	Midterm Project Report
10	2025/11/11	NVIDIA Fundamentals of Deep Learning III: Pre-trained Models; Natural Language Processing

Syllabus

Week	Date	Subject/Topics
11	2025/11/18	Case Study on Artificial Intelligence II
12	2025/11/25	Computer Vision and Robotics
13	2025/12/02	Generative AI, Agentic AI, and Physical AI
14	2025/12/09	Philosophy and Ethics of AI and the Future of AI
15	2025/12/16	Final Project Report I
16	2025/12/23	Final Project Report II

Philosophy and Ethics of AI and the Future of AI

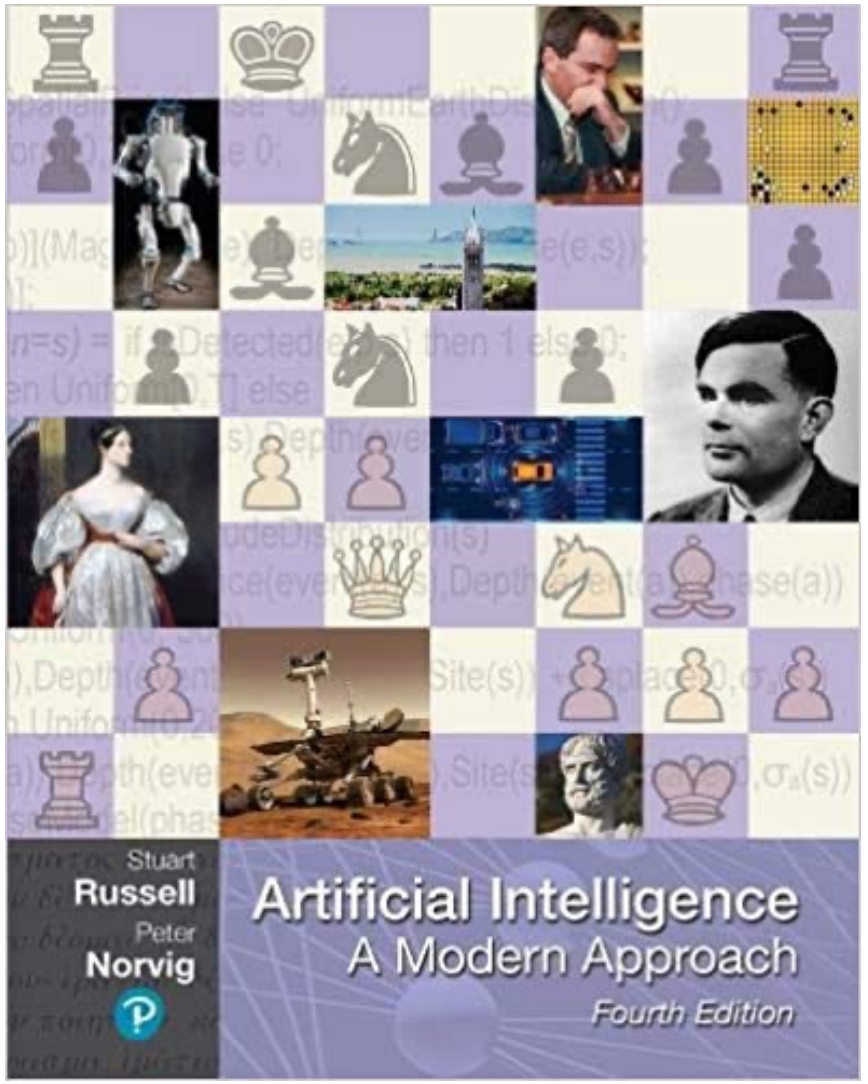
Outline

- Philosophy, Ethics, and Safety of AI
  - The Limits of AI
  - Can Machines Really Think?
  - The Ethics of AI
- The Future of AI
  - AI Components
  - AI Architectures

Stuart Russell and Peter Norvig (2020),

Artificial Intelligence: A Modern Approach,

4th Edition, Pearson



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

https://www.amazon.com/Artificial-Intelligence-A-Modern-Approach/dp/0134610997/

Artificial Intelligence: A Modern Approach

1. Artificial Intelligence
2. Problem Solving
3. Knowledge and Reasoning
4. Uncertain Knowledge and Reasoning
5. Machine Learning
6. Communicating, Perceiving, and Acting
7. Philosophy and Ethics of AI

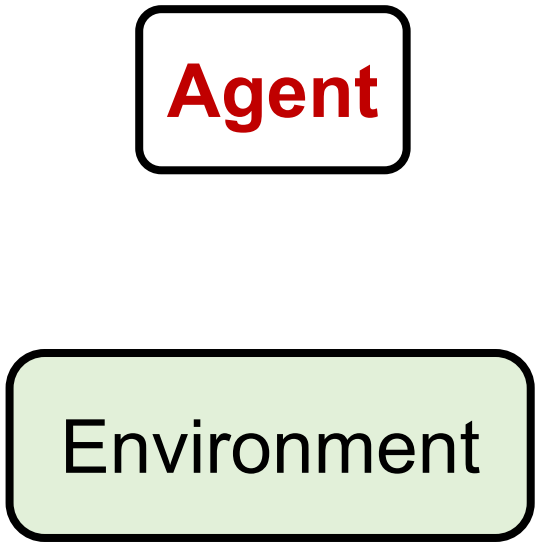
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Artificial Intelligence:
7. Philosophy and Ethics of AI

- Philosophy, Ethics, and Safety of AI
  - The Limits of AI
  - Can Machines Really Think?
  - The Ethics of AI
- The Future of AI
  - AI Components
  - AI Architectures

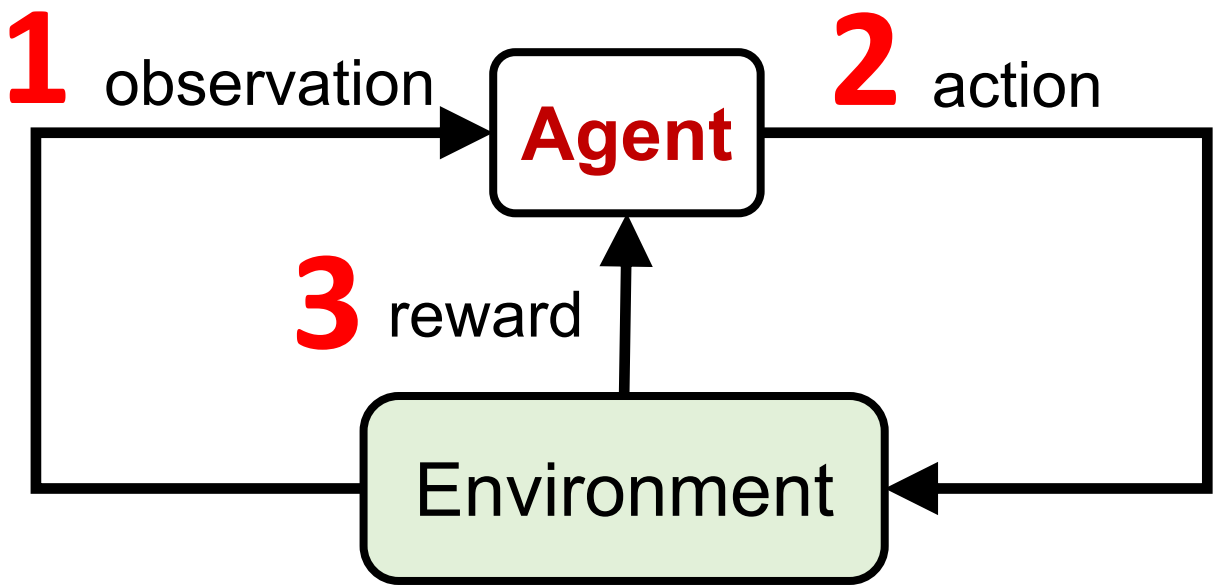
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Reinforcement Learning (DL)



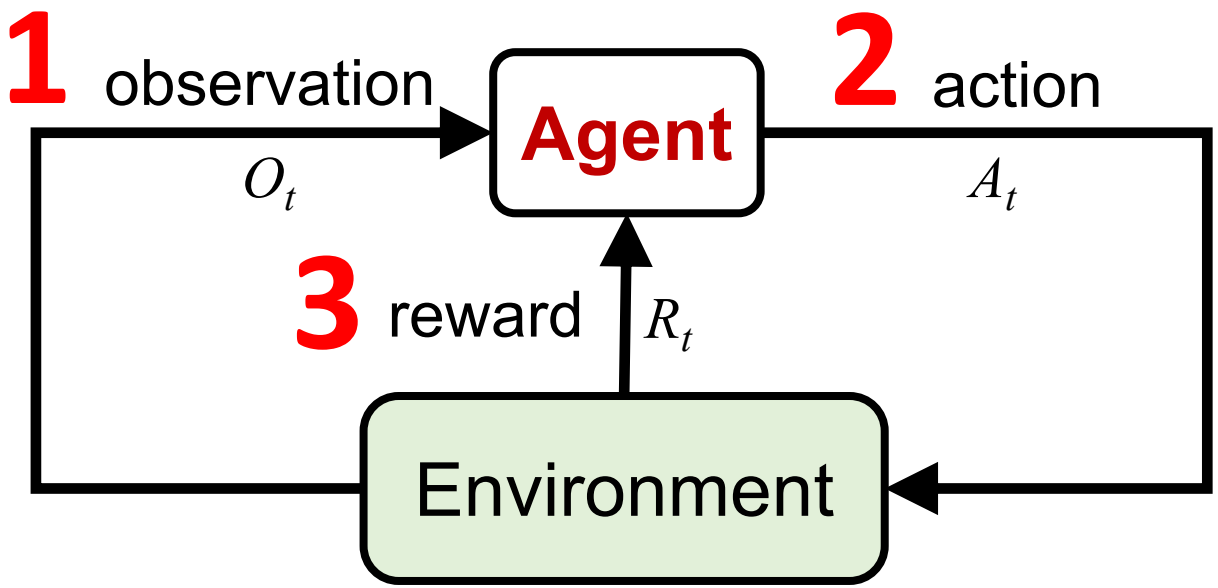
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



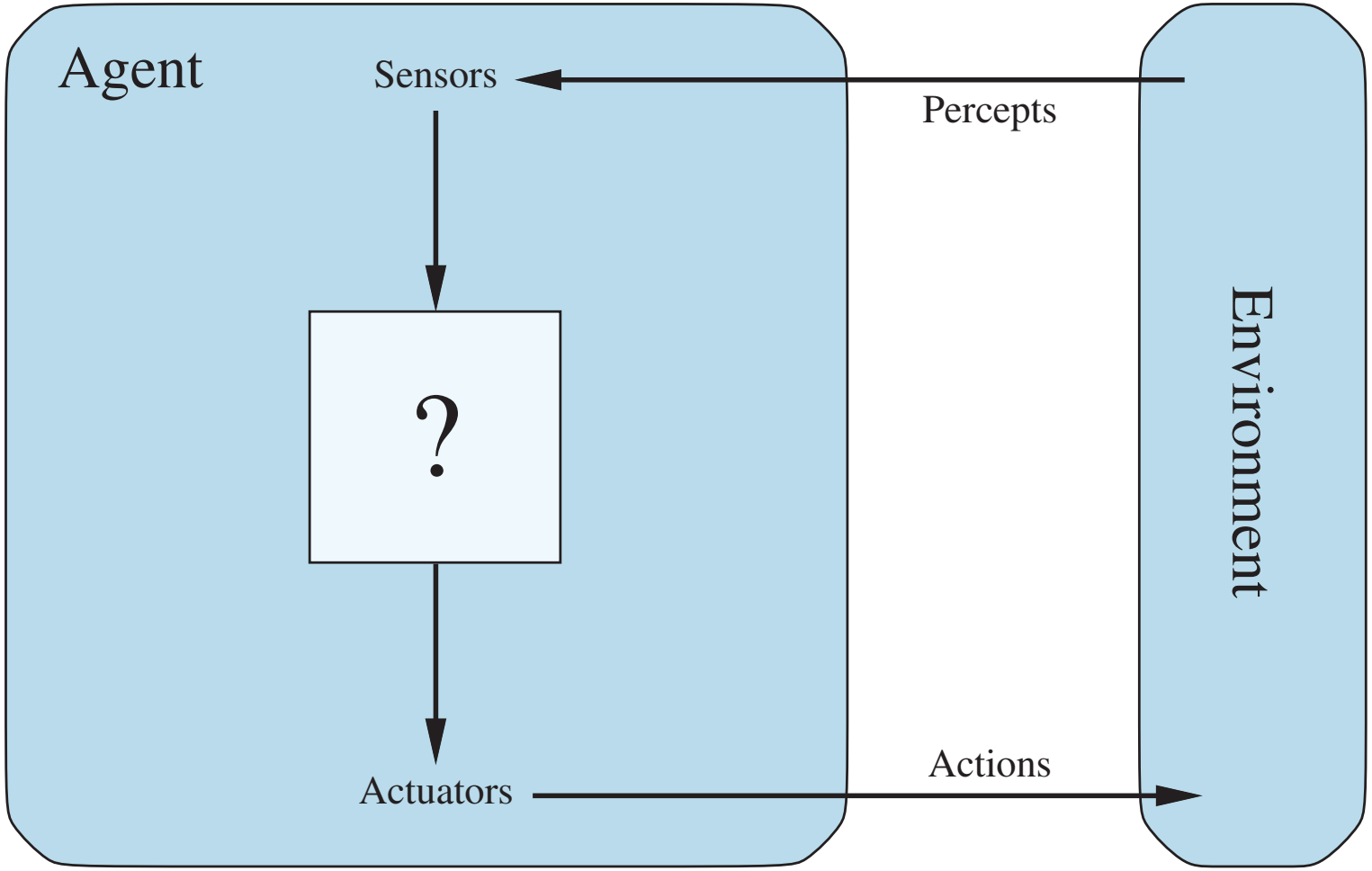
Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Reinforcement Learning (DL)



Source: Richard S. Sutton & Andrew G. Barto (2018), Reinforcement Learning: An Introduction, 2nd Edition, A Bradford Book.

Agents interact with environments through sensors and actuators



Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

AI Acting Humanly:
The Turing Test Approach
(Alan Turing, 1950)

- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
  - Deep Learning (DL)
- Computer Vision (Image, Video)
- Natural Language Processing (NLP)
- Robotics

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Philosophy and Ethics of AI

Philosophy, Ethics, and Safety of AI

- The Limits of AI
- Can Machines Really Think?
- The Ethics of AI

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Philosophy of AI

- Philosophers use the term
  - **weak AI** for the hypothesis that machines could possibly behave intelligently
  - **strong AI** for the hypothesis that such machines would count as having actual minds (as opposed to simulated minds)

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

4 Approaches of AI

Thinking Humanly	**Thinking Rationally**
Acting Humanly	**Acting Rationally**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

4 Approaches of AI

2. **Thinking Humanly:** **The Cognitive Modeling Approach**	**3.** **Thinking Rationally:** **The "Laws of Thought" Approach**
1. **Acting Humanly:** **The Turing Test Approach (1950)**	**4.** **Acting Rationally:** **The Rational Agent Approach**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

AI Acting Humanly:
The Turing Test Approach
(Alan Turing, 1950)

- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
  - Deep Learning (DL)
- Computer Vision (Image, Video)
- Natural Language Processing (NLP)
- Robotics

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Can machines think?

- Alan Turing rejected the question “Can machines think?” and replaced it with a behavioral test.
  - Alan Turing anticipated many objections to the possibility of thinking machines.
- Concentrate on their systems’ performance on practical tasks
  - rather than the ability to imitate humans.
- Consciousness remains a mystery.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

The Ethics of AI

- **Given that AI is a powerful technology, we have a moral obligation to use it well, to promote the positive aspects and avoid or mitigate the negative ones.**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Principles of Robotics and AI

- Ensure safety
- Ensure fairness
- Respect privacy
- Promote collaboration
- Provide transparency
- Limit harmful uses of AI

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Principles of Robotics and AI

- Establish accountability
- Uphold human rights and values
- Reflect diversity/inclusion
- Avoid concentration of power
- Acknowledge legal/policy implications
- Contemplate implications for employment

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Safety of AI

- AI is a powerful technology, and as such it poses potential dangers, through lethal autonomous weapons, security and privacy breaches, unintended side effects, unintentional errors, and malignant misuse.
- Those who work with AI technology have an ethical imperative to responsibly reduce those dangers.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Robot Ethics

Laws of Robotics (Isaac Asimov, 1942, 1950)

1. **A robot may not injure a human being or, through inaction, allow a human being to come to harm.**
2. **A robot must obey orders given to it by human beings, except where such orders would conflict with the First Law.**
3. **A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.**

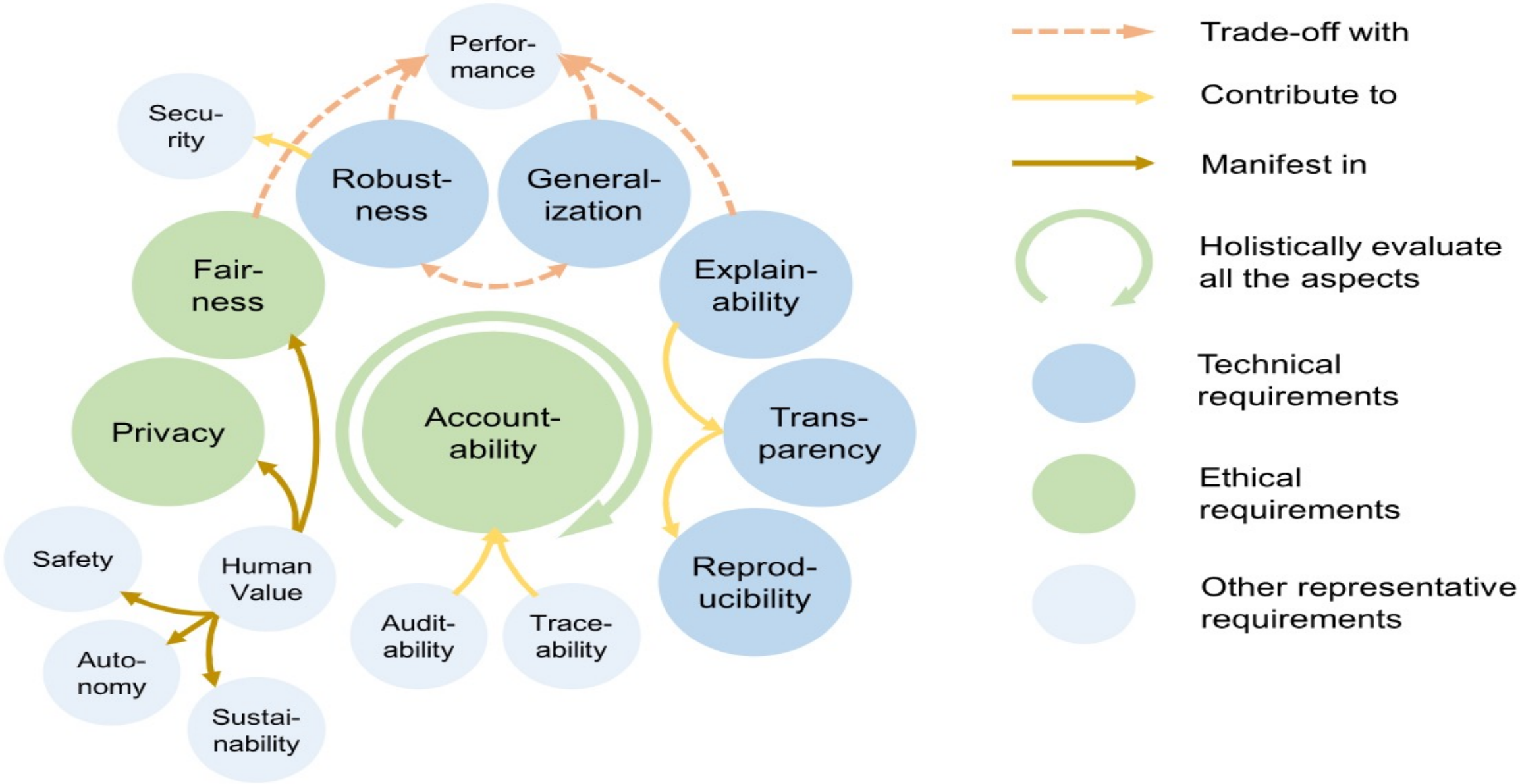
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Fair, trustworthy, and transparent of AI

- AI systems must be able to demonstrate they are fair, trustworthy, and transparent.
- There are multiple aspects of fairness, and it is impossible to maximize all of them at once.
- So a first step is to decide what counts as fair.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Trustworthy AI



Source: Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. (2023) "Trustworthy AI: From principles to practices." ACM Computing Surveys 55, no. 9 (2023): 1-46.

Explainable AI (XAI)

- When an AI system turns you down for a loan, you deserve an explanation.
  - In Europe, the GDPR enforces this for you.
  - An AI system that can explain itself is called **explainable AI (XAI)**.

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Explainable AI (XAI)

- A good explanation properties of XAI
  - it should be understandable and convincing to the user
  - it should accurately reflect the reasoning of the system
  - it should be complete
  - it should be specific in that different users with different conditions or different outcomes should get different explanations

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

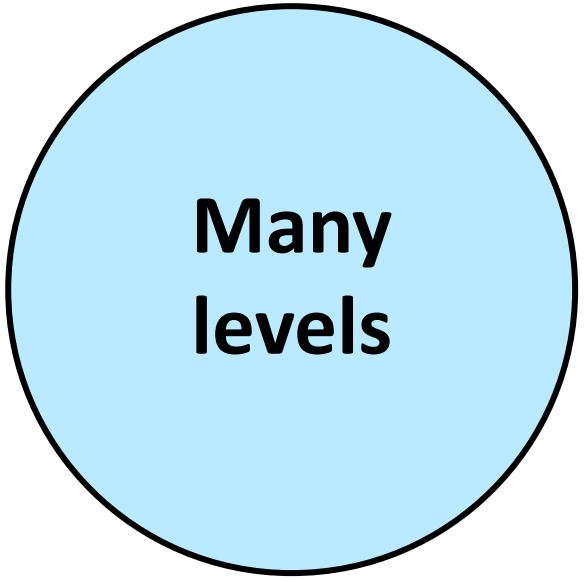
Automation

- **Automation is already changing the way people work.**
- **As a society, we will have to deal with these changes.**

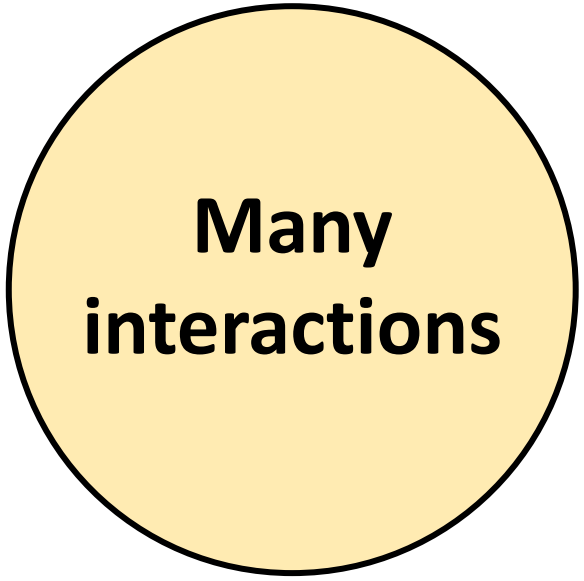
M3 Framework for AI Responsibility

Many agents, Many levels, and Many interactions

Multiple individuals and organizations contribute to LLM outcomes



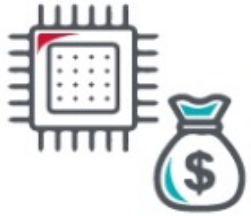
Responsibility spans individual, organizational, regulatory, and global layers



Influence across actors accumulates and grounds responsibility

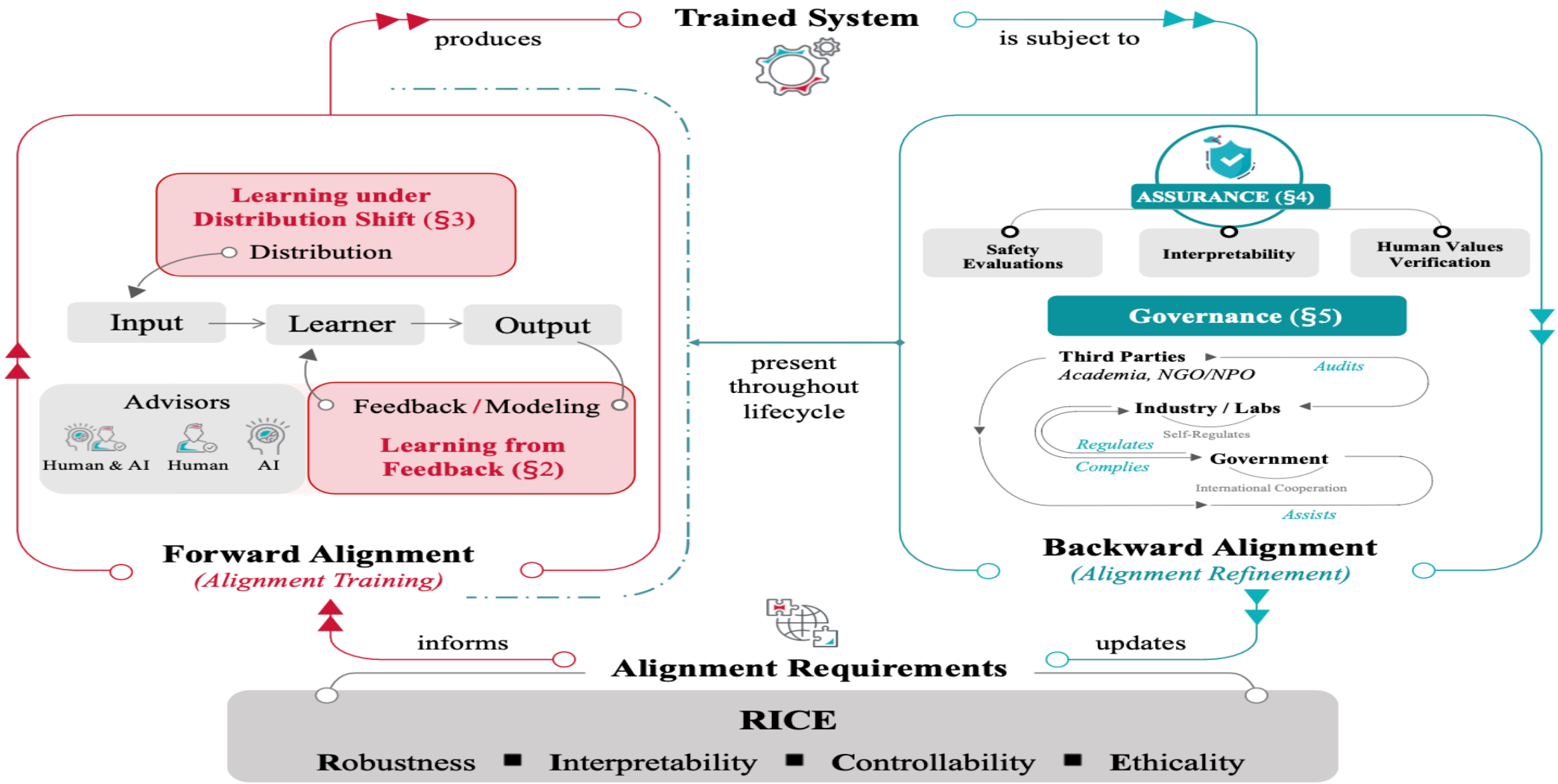
Source: Mihaela Constantinescu, and Muel Kaptein. (2025) "Responsibility gaps, LLMs & organisations: Many agents, many levels, and many interactions." Science and Engineering Ethics 31, no. 6 (2025): 36.

AI Alignment:
Dangerous Capabilities
Arising from AI Power-Seeking

 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

Source: Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys.

The AI Alignment Cycle: Training and Refinement



Source: Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys.

The RICE principles of AI Alignment:
Robustness, Interpretability, Controllability, and Ethicality

Robustness — Operates reliably under diverse scenarios | Resilient to unforeseen disruptions.

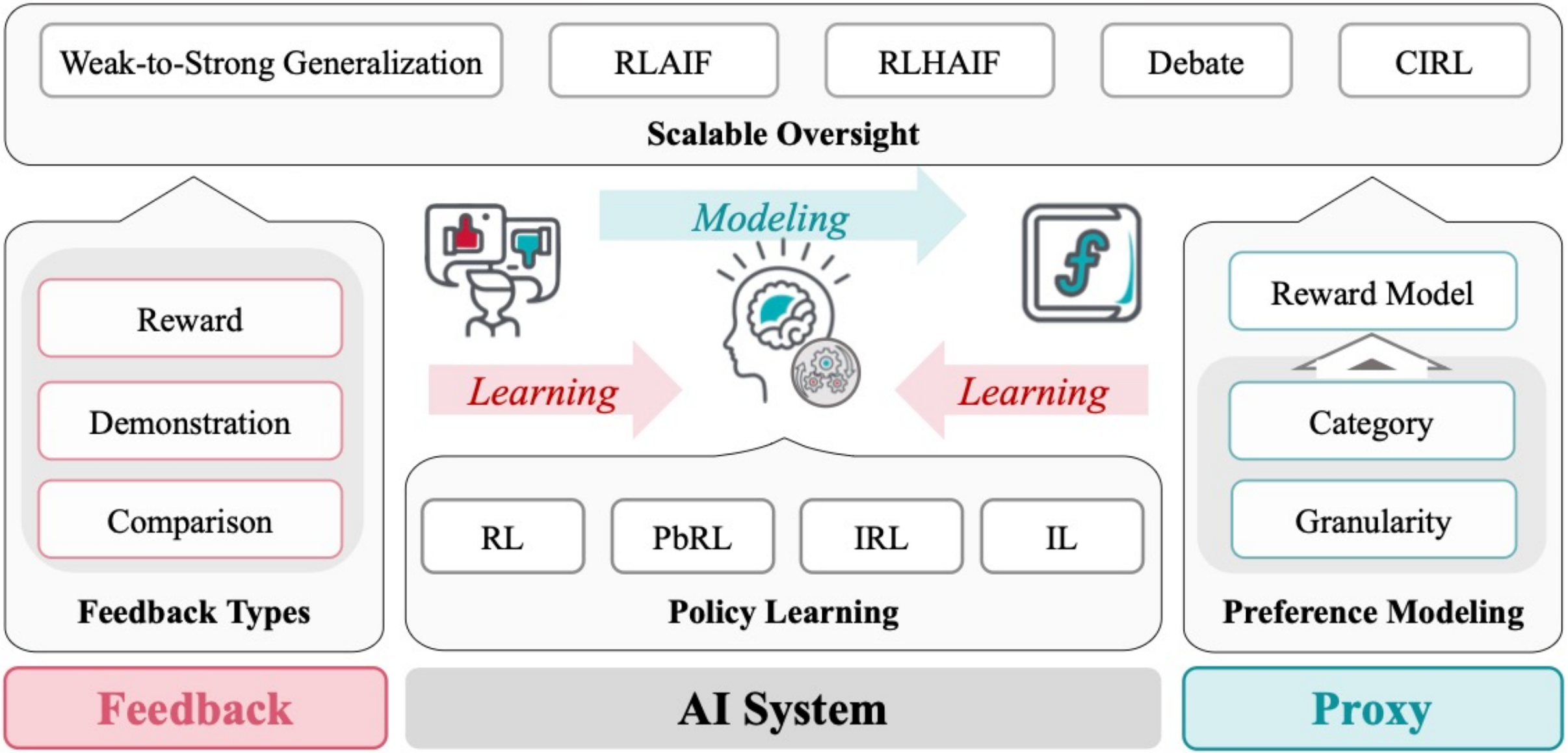
Interpretability — Decisions and intentions are comprehensible | Reasoning is unconcealed and truthful.

Controllability — Behaviors can be directed by humans | Allows human intervention when needed.

Ethicality — Adheres to global moral standards | Respects values within human society.

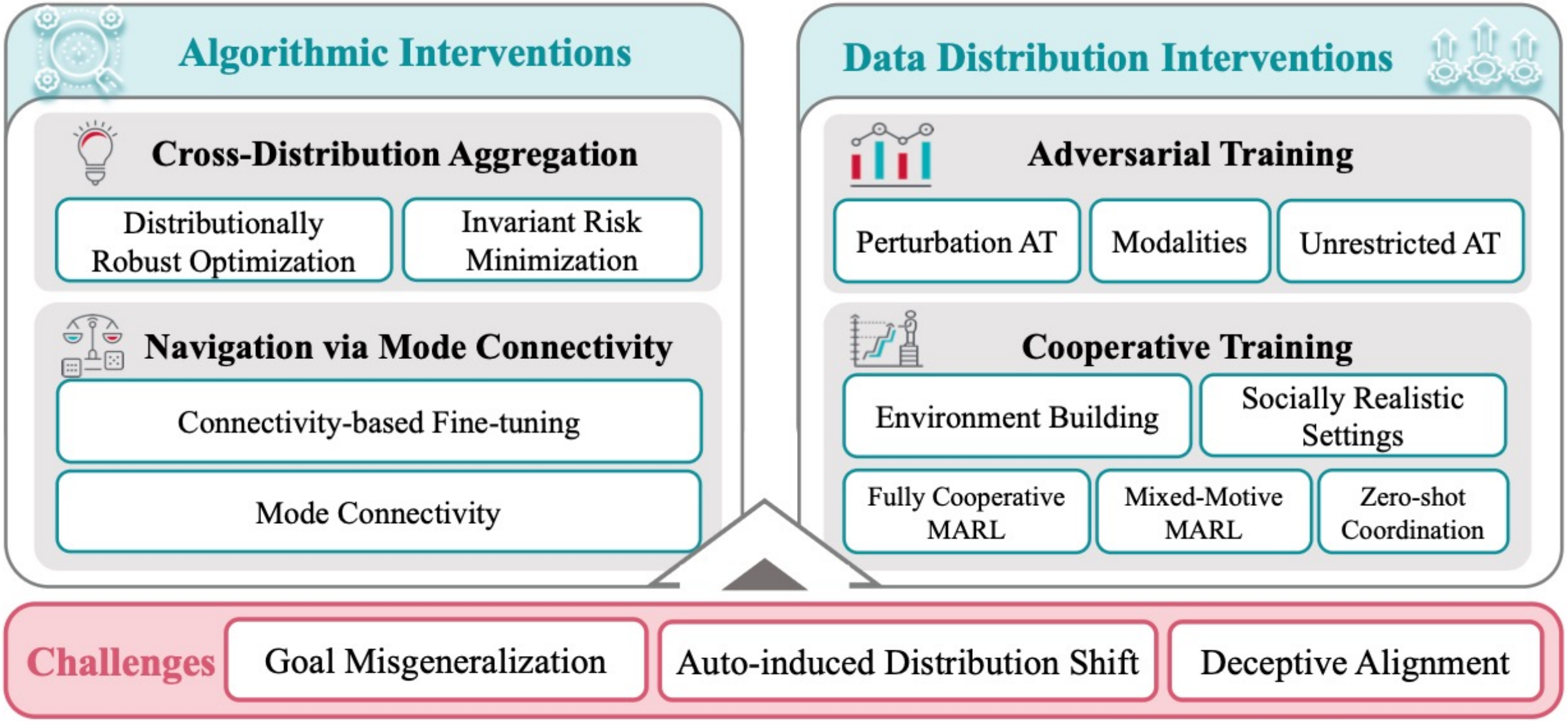
Source: Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys.

Learning from Human Feedback in AI Alignment



Source: Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys.

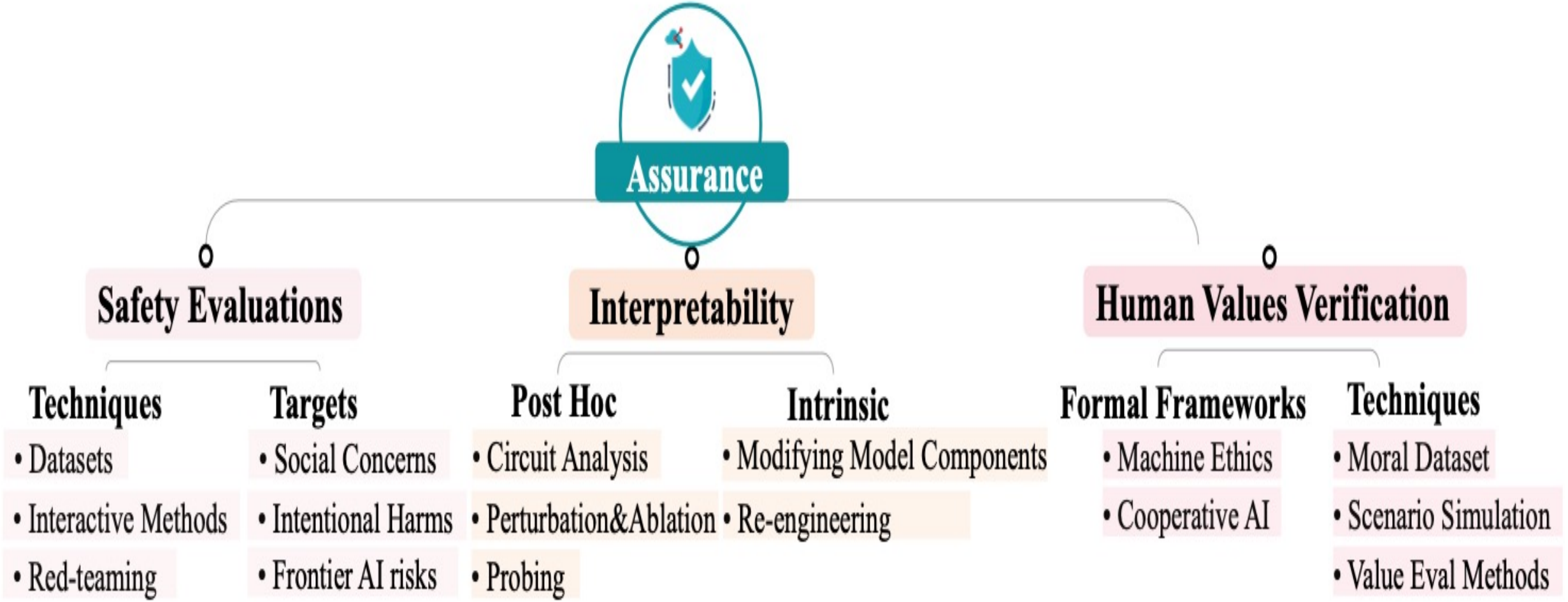
Learning Under Distribution Shift in AI Alignment



Source: Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys.

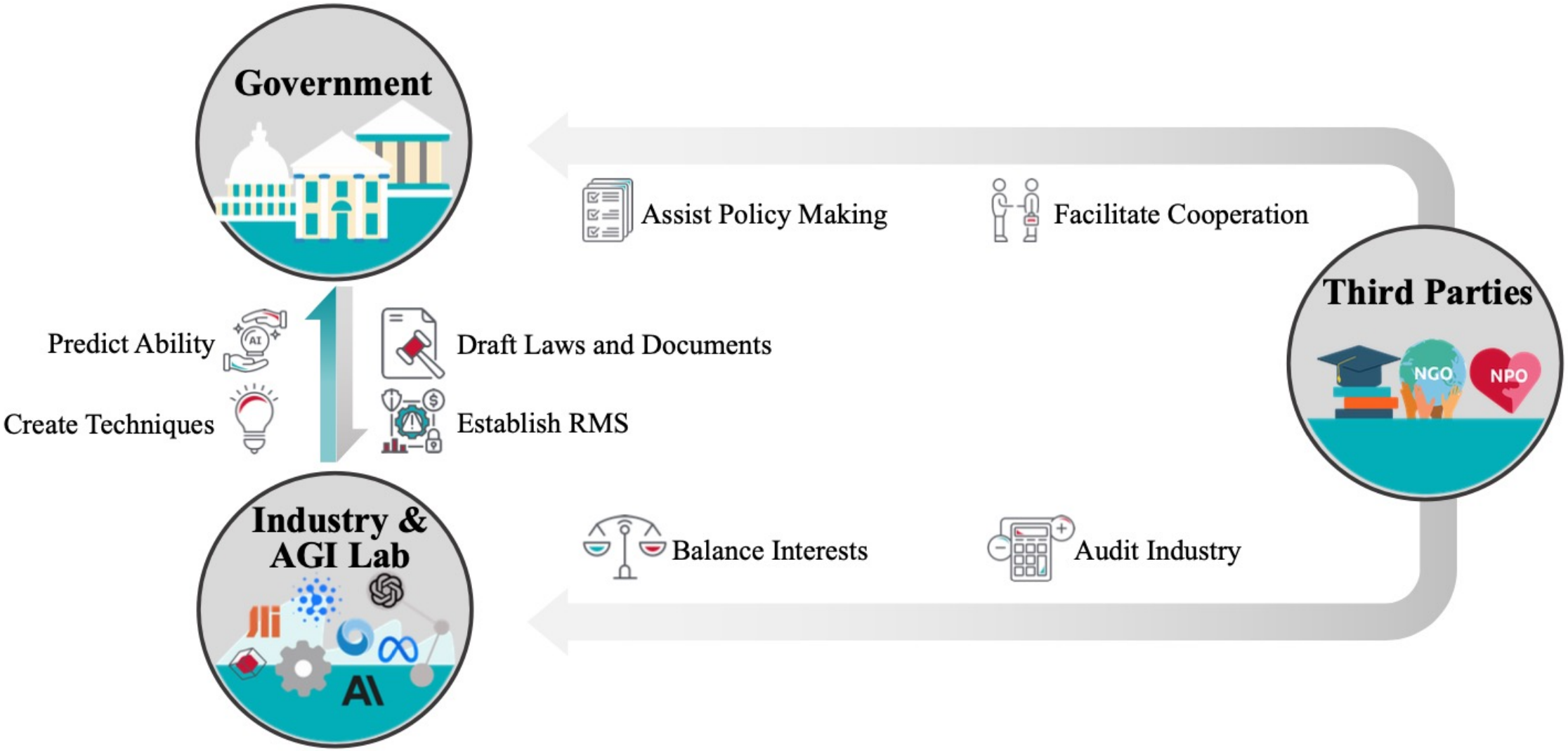
Research Directions in AI Alignment:

Safety Evaluation, Interpretability, and Human Value Verification



Source: Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys.

AI Governance Framework: Government, Industry & AGI Lab, and Third Parties



Source: Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys.

The Future of AI

Generative AI, Agentic AI, Physical AI

2012 AlexNet
Deep learning breakthrough

Perception AI
Speech recognition
Deep recommender systems
Medical imaging

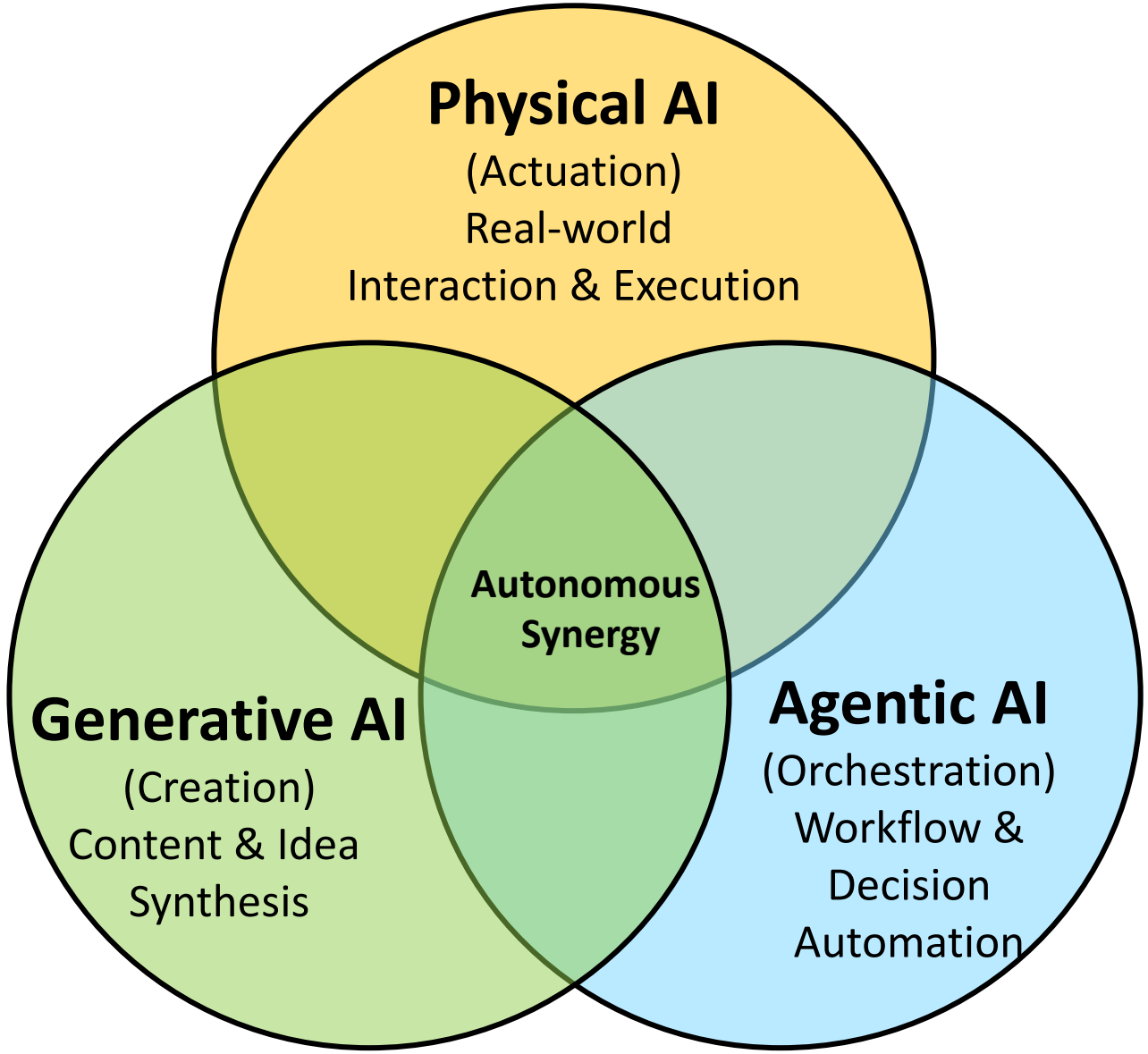
Generative AI
Digital marketing
Content creation

Agentic AI
Coding assistants
Customer service
Patient care

Physical AI
Self-driving cars
General robotics

Source: NVIDIA (2025), GTC March 2025 Keynote with NVIDIA CEO Jensen Huang, https://www.youtube.com/watch?v=_waPvOwL9Z8

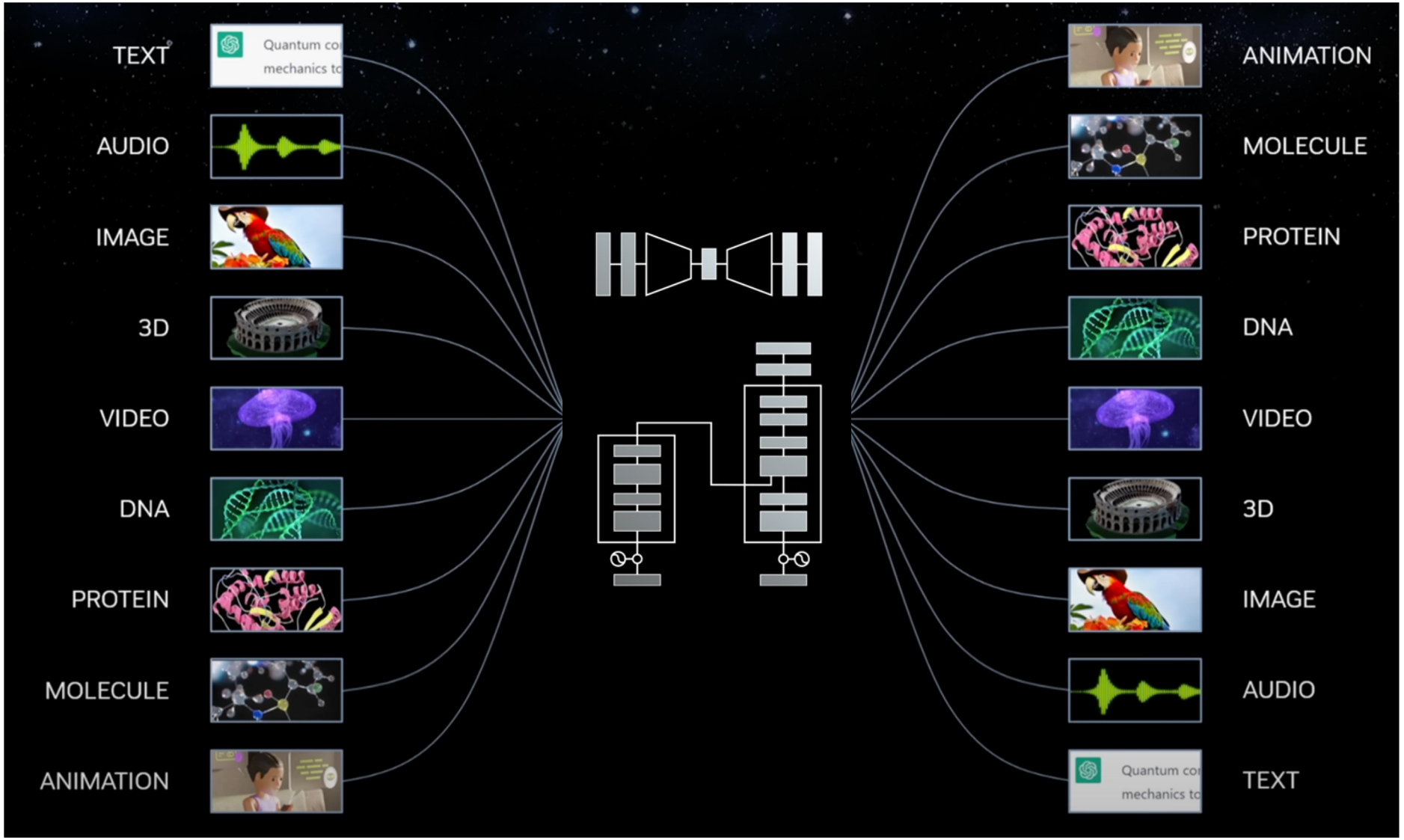
Generative AI, Agentic AI, Physical AI



New Economic Paradigm Shift: From Creation to Execution

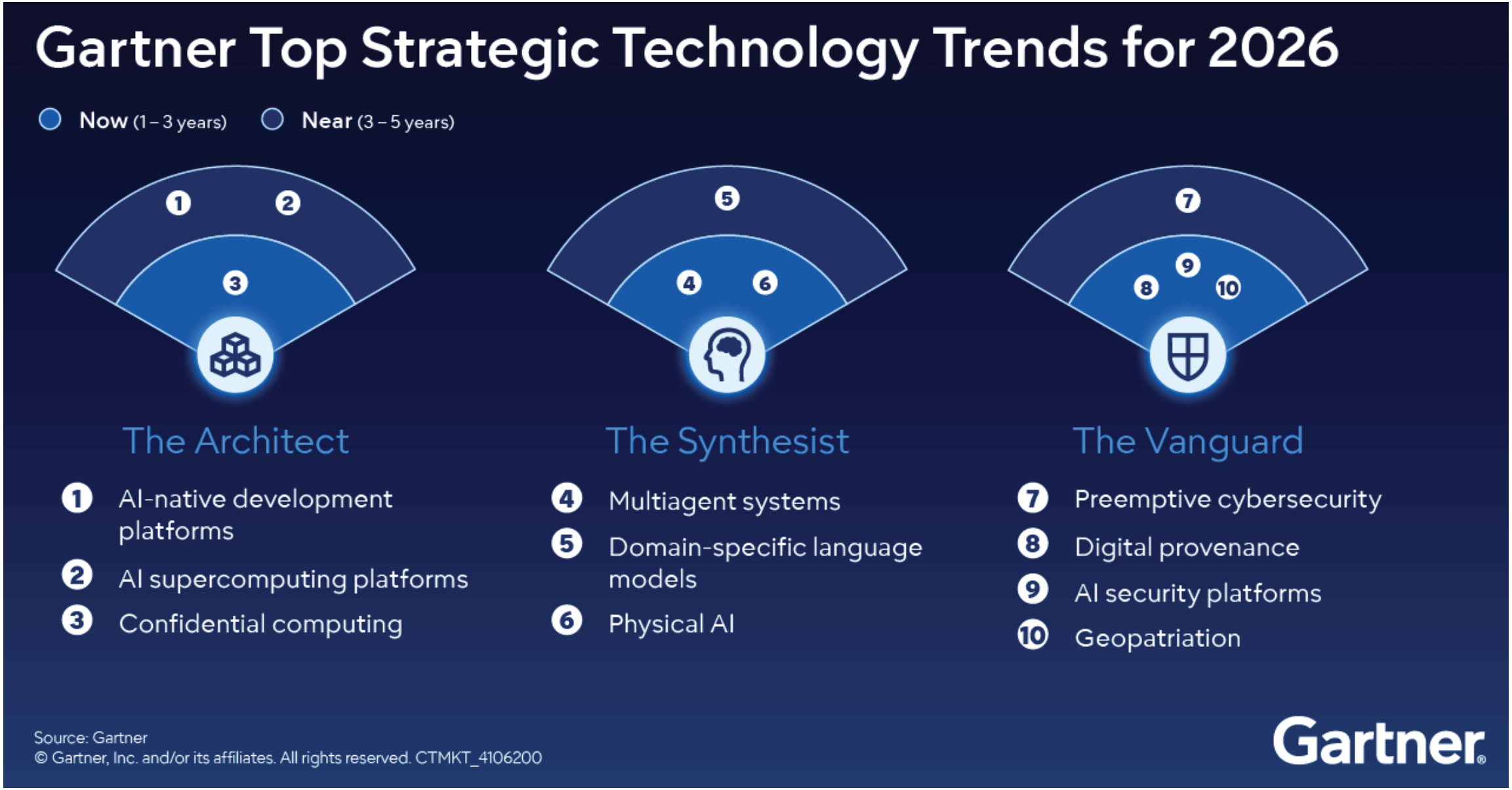
Modular Modalities

Where Can The Transformer Fit?



Source: NVIDIA DLI (2025), Rapid Application Development with Large Language Models (LLMs), https://learn.nvidia.com/courses/course-detail?course_id=course-v1:DLI+S-FX-26+V1

Gartner Top 10 Strategic Technology Trends for 2026



Source: https://www.gartner.com/en/articles/top-technology-trends-2026

The Architect: AI platforms and infrastructure

building secure, scalable foundations for AI and digital transformation

- **AI-Native Development Platforms**
  - empower small, nimble teams to build software using generative AI fast, flexible and increasingly enterprise-ready.
- **AI Supercomputing Platforms**
  - unlock breakthroughs in model training and analytics, but require careful governance and cost control.
- **Confidential Computing**
  - protects sensitive data while in use, enabling secure AI and analytics across untrusted infrastructure.

Source: https://www.gartner.com/en/articles/top-technology-trends-2026

The Synthesist: AI application and orchestration

combine specialized models, agents and physical-digital systems to create new value

- **Multiagent Systems (MAS)**
  - allow modular AI agents to collaborate on complex tasks, improving automation and scalability.
- **Domain-Specific Language Models (DSLM)**
  - deliver higher accuracy and compliance for industry-specific use cases.
- **Physical AI (Robotics)**
  - brings intelligence into the real world
    powering robots, drones and smart equipment for operational impact.

Source: https://www.gartner.com/en/articles/top-technology-trends-2026

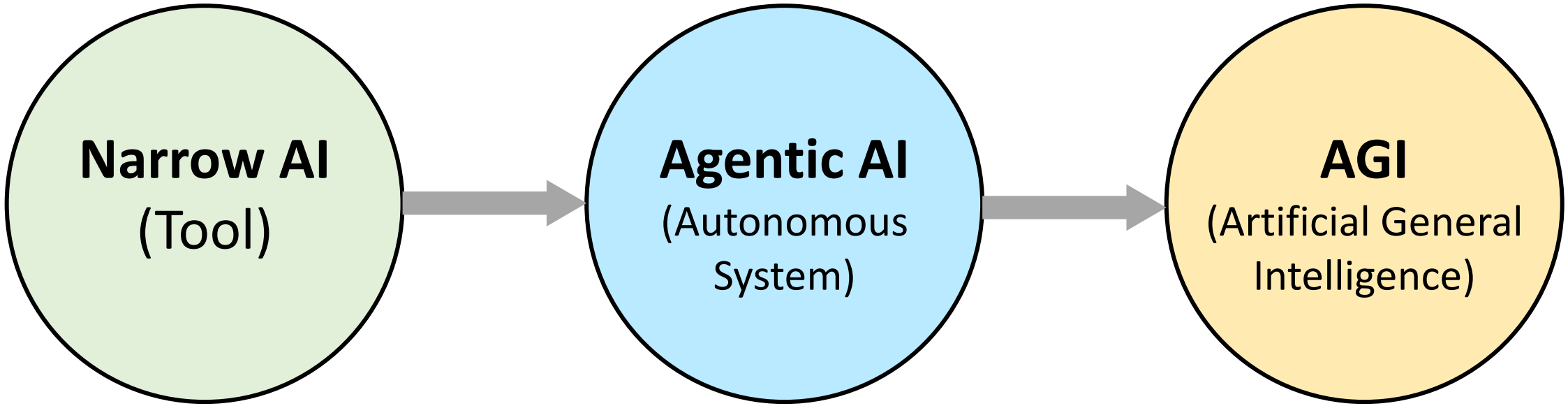
The Vanguard: security, trust and governance

protect reputation, ensure compliance and maintain stakeholder confidence

- **Preemptive Cybersecurity**
  - shifts defense from reactive to proactive, using AI to block threats before they strike
- **Digital Provenance**
  - verifies the origin and integrity of software, data and AI-generated content essential for trust and compliance
- **AI Security Platforms**
  - centralize visibility and control across third-party and custom AI applications
- **Geopatriation**
  - helps organizations mitigate geopolitical risk by shifting workloads to sovereign or regional cloud providers

Source: https://www.gartner.com/en/articles/top-technology-trends-2026

The Future of AI
From Tools to Agents:
The Rise and Autonomy of Agentic AI



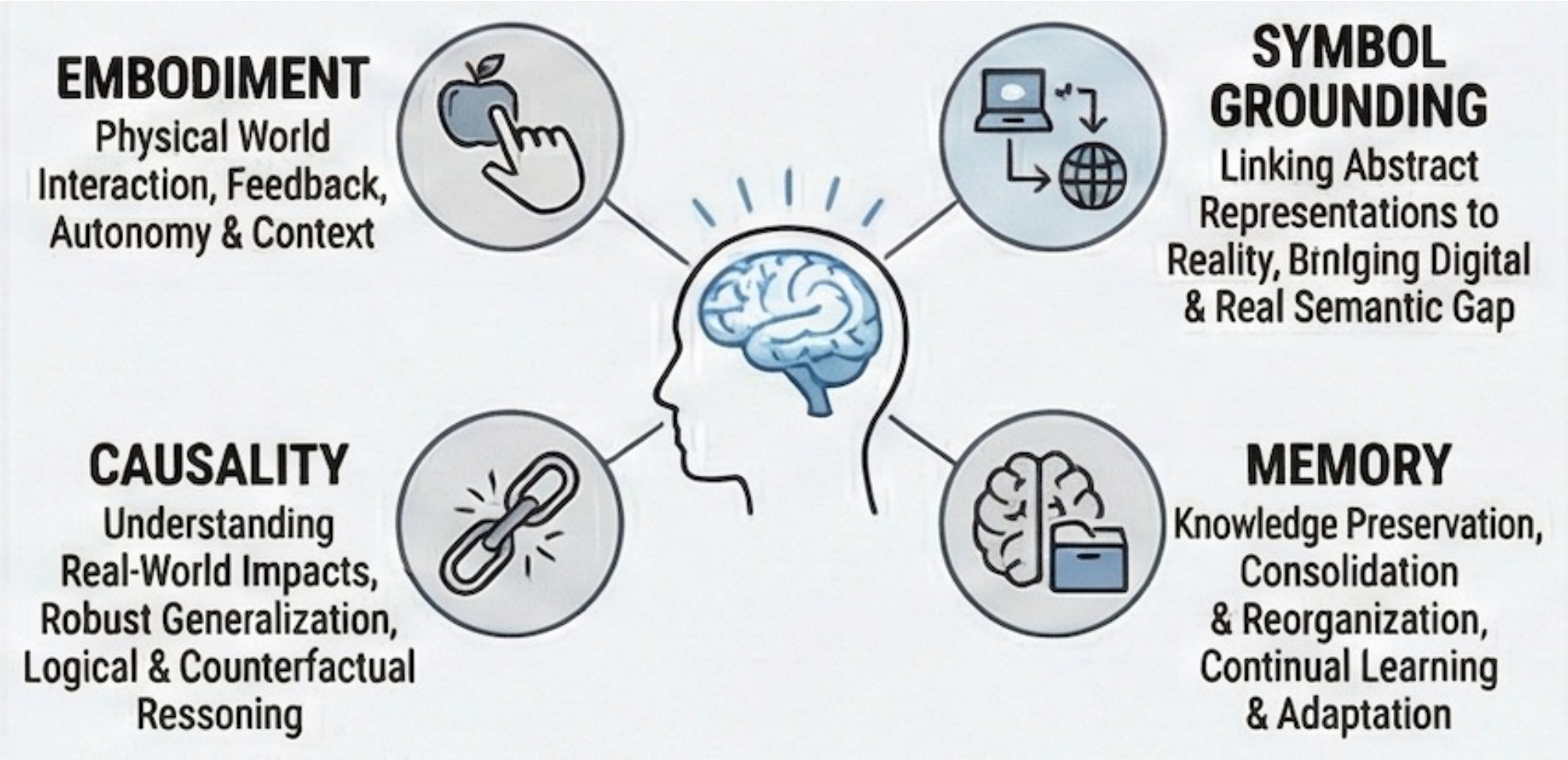
Executes Specific Tasks,
Content Generation

Goal-Oriented Behavior,
Autonomous Decision-Making,
Multi-Step Workflows,
Self-Improvement within Parameters

Connecting Multimodal GenAI
to Future AGI

The Future of AI

Deep Integration of Core Cognitive Models: Toward Embodied AGI



The Future of AI

- AI Components
- AI Architectures

AI Components

- Sensors and actuators
- Representing the state of the world
- Selecting actions
- Deciding what we want
- Learning
- Resources
  - Shared data
  - Shared model

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

Learning

- Deep learning
- Data science
- Big data
- Transfer learning
- Apprenticeship learning
- Differentiable programming
- Weakly supervised learning
- Predictive learning

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

AI Architectures

- Which of the agent architectures should an agent use?
  - All of them!
- Real-time AI
- Anytime algorithm
- Decision-theoretic metareasoning
- Reflective architecture
- Agent = Architecture + Program
- Bounded optimality

Real-time AI

- **As AI systems move into more complex domains, all problems will become real-time, because the agent will never have long enough to solve the decision problem exactly.**

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

General AI

- Narrow tasks AI
  - DARPA Grand Challenge for autonomous cars
  - ImageNet object recognition competition
  - For each separate task, we build a separate AI system
  - A separate machine learning model trained from scratch with data collected specifically for this task.
- Human-level AI (HLAI)

Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

AI Engineering

- Powerful tools and frameworks
  - TensorFlow, Keras, PyTorch, CAFFE, Scikit-Learn and SCIPY.
- Promising approaches
  - GANs
  - Deep reinforcement learning
  - Train properly in a new domain

The Future of AI

- AI has made great progress in its short history.
- We can see only a short distance ahead, but we can see that much remains to be done.

  (Alan Turing, 1950)

  [Computing Machinery and Intelligence]

The Future of AI

- Past: Build each new system from scratch
- Future: Start with a single huge system
  - For each new task, extract from it the parts that are relevant to the task.
- Transformer language models
  (e.g., BERT, GPT-3, ChatGPT, Claude, Gemini)

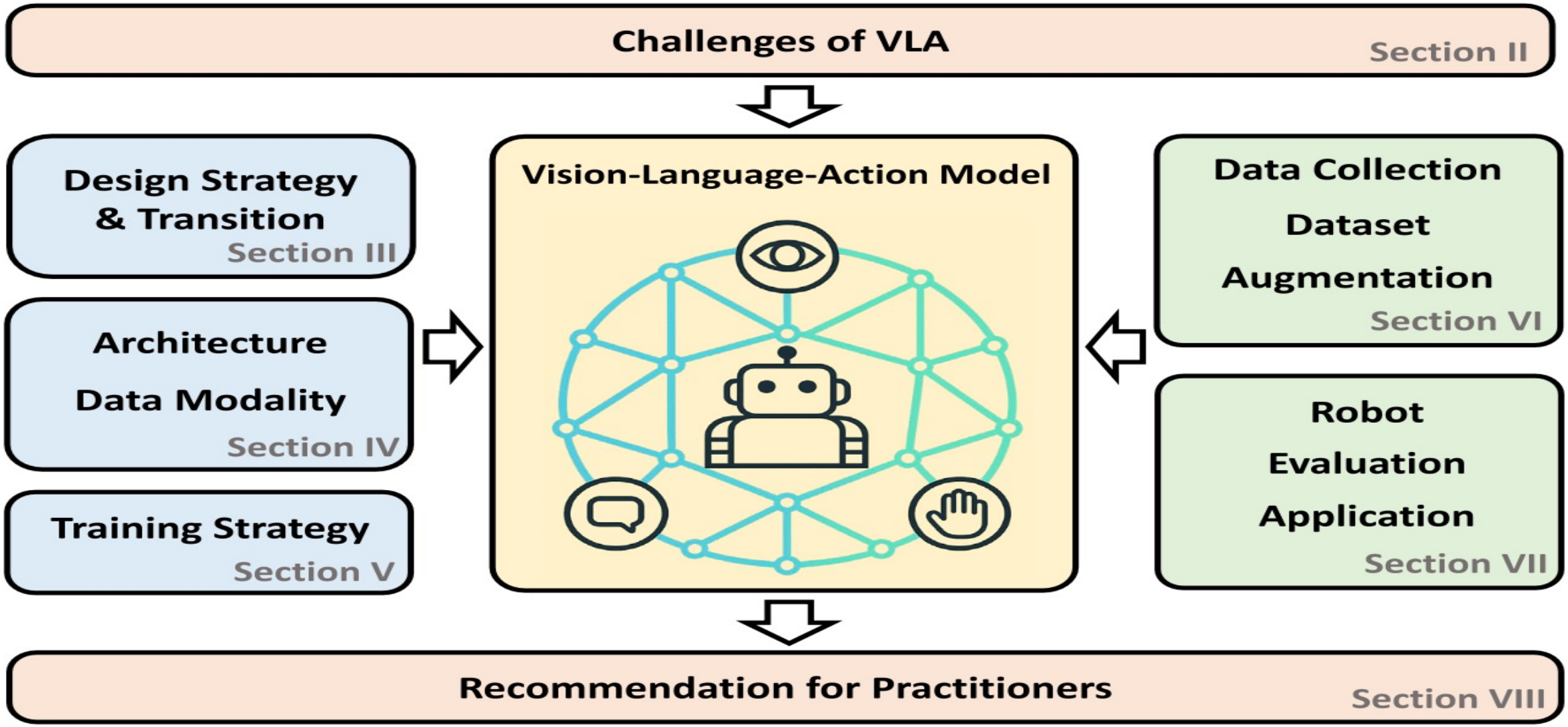
Source: Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson

The Future of AI

- **From Tools to Agents:**
  The Rise and Autonomy of Agentic AI
- **Deep Integration of Core Cognitive Models:**
  Toward Embodied AGI
- **Governance Shifts and the**
  Restructuring of Socio-Economic Systems

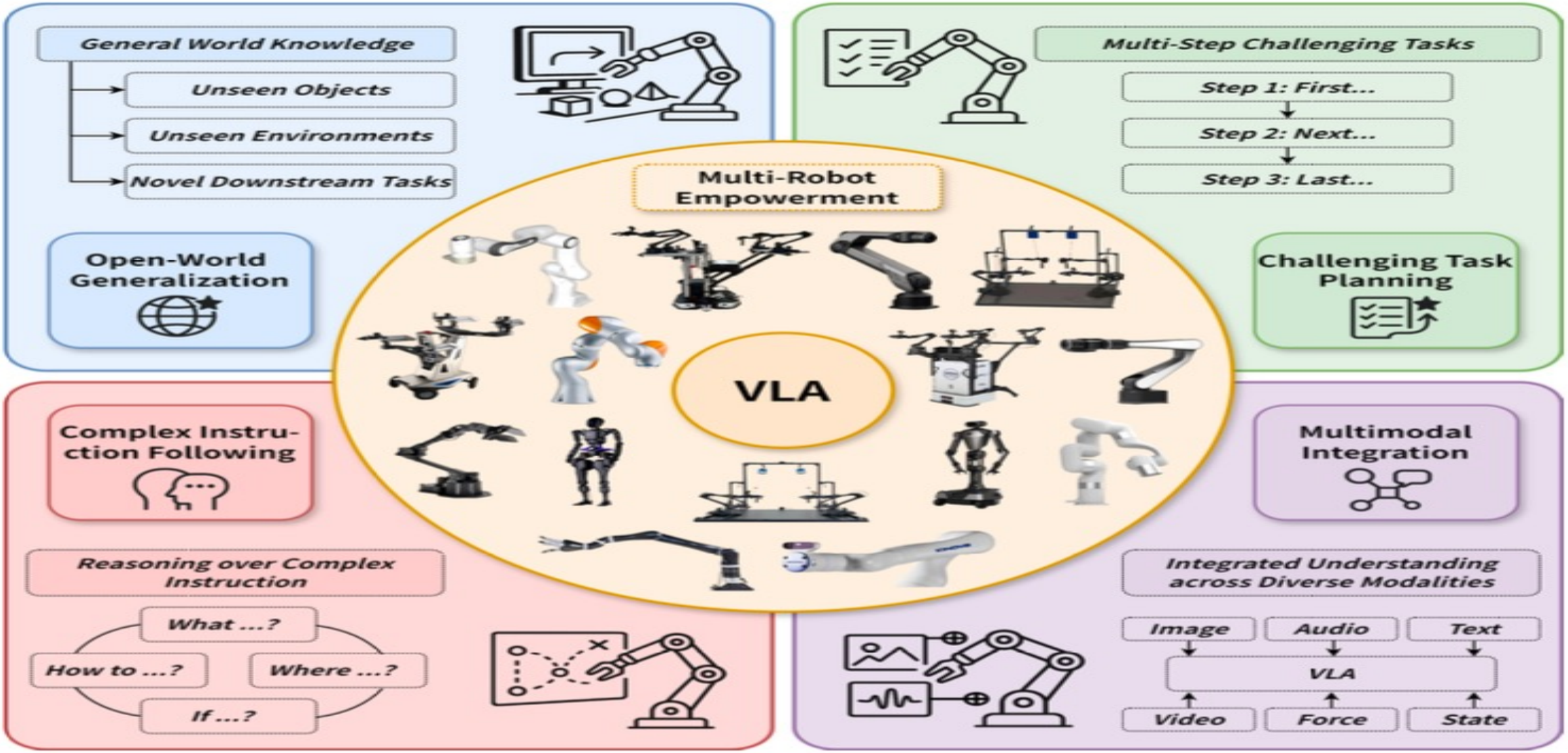
Vision Language Action (VLA) Models for Robotics

Vision-Language-Action (VLA) Models for Robotics



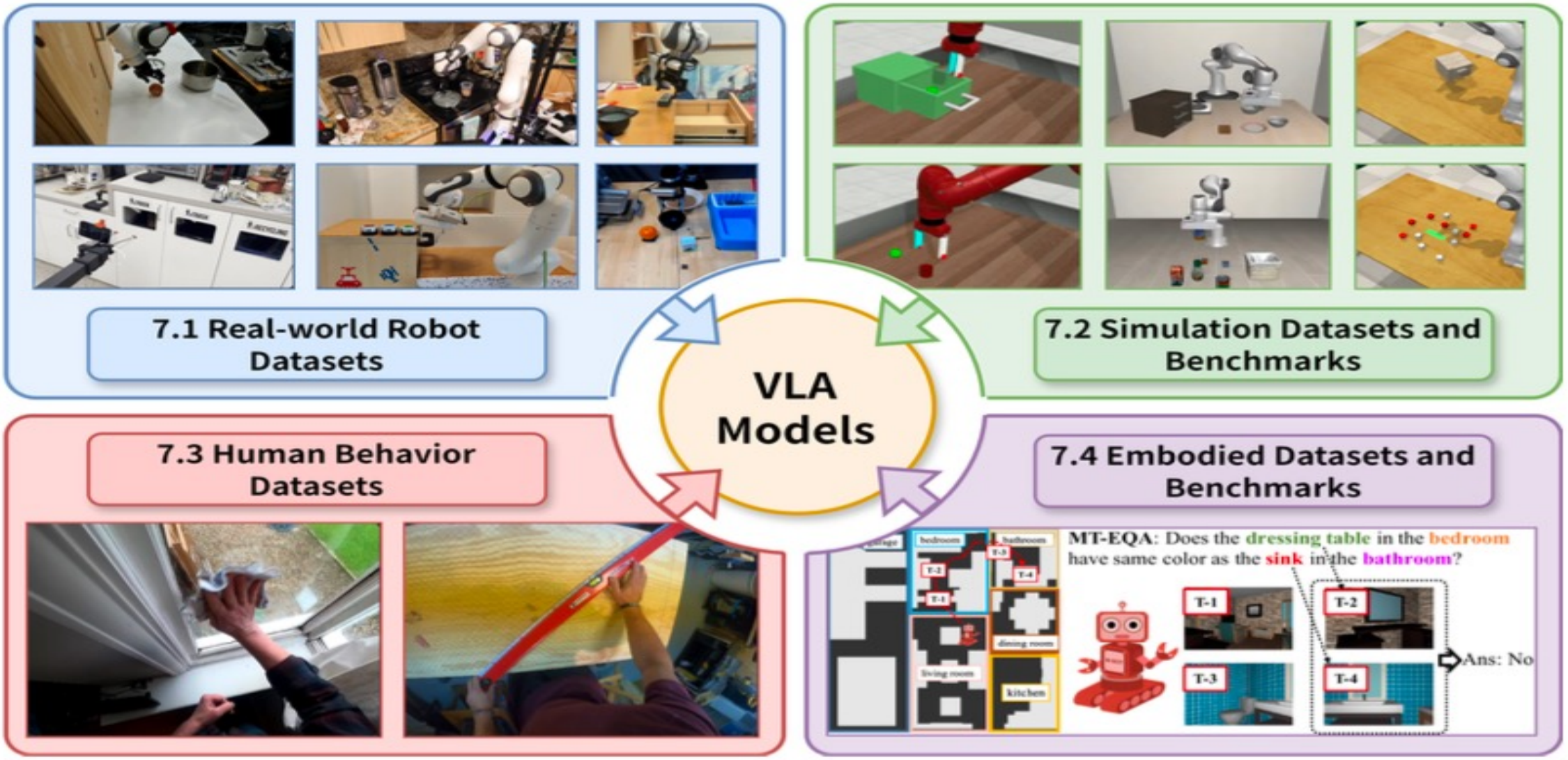
Source: Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).

Large VLM-based Vision-Language-Action Models for Robotic Manipulation



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

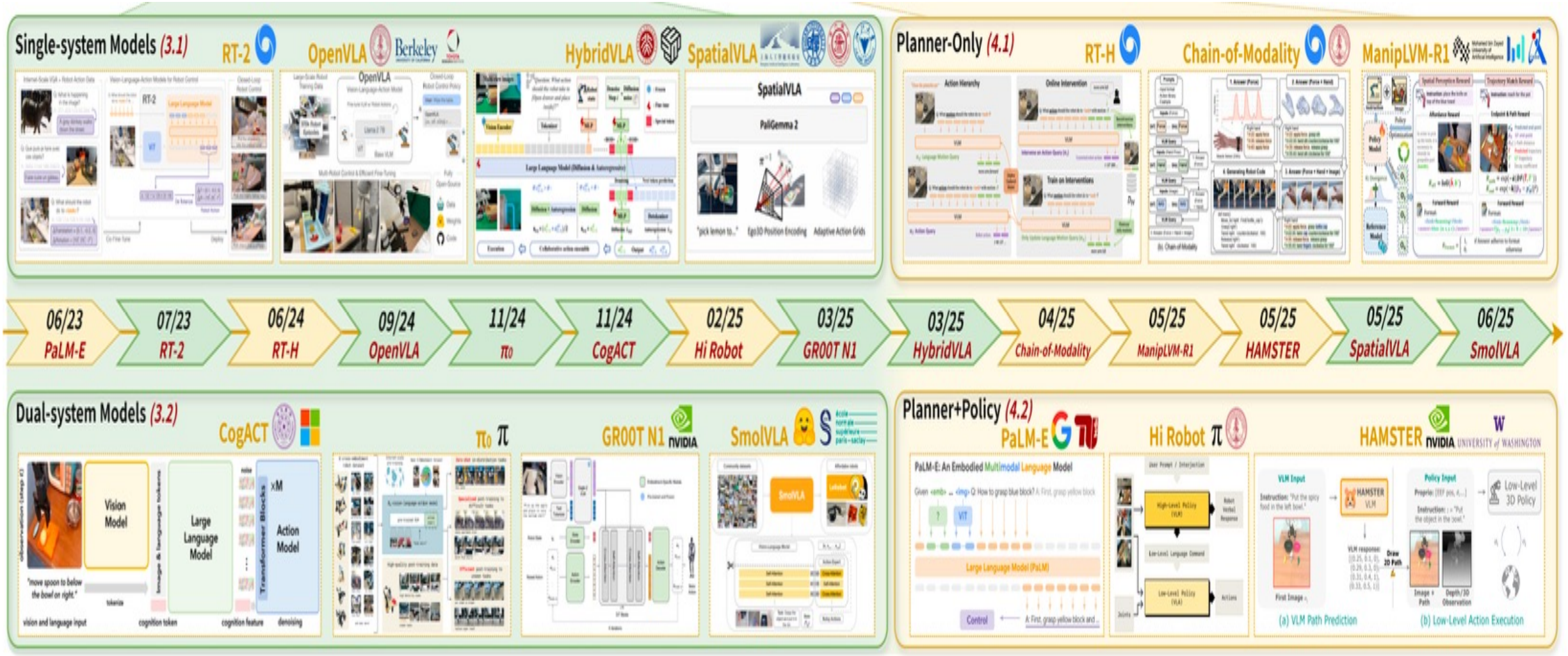
Large VLM-based Vision-Language-Action Models for Robotic Manipulation



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

Large VLM-based Vision-Language-Action Models for Robotic Manipulation (Timeline)

Monolithic models and Hierarchical Models



Source: Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).

AI for Social Good (AI4SG)

Source: Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020) "AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.

Sustainable Development Goals (SDGs)

Source: https://sdgs.un.org/goals

Sustainable Development Goals (SDGs) and 5P

Partnership

Peace

Prosperity

People

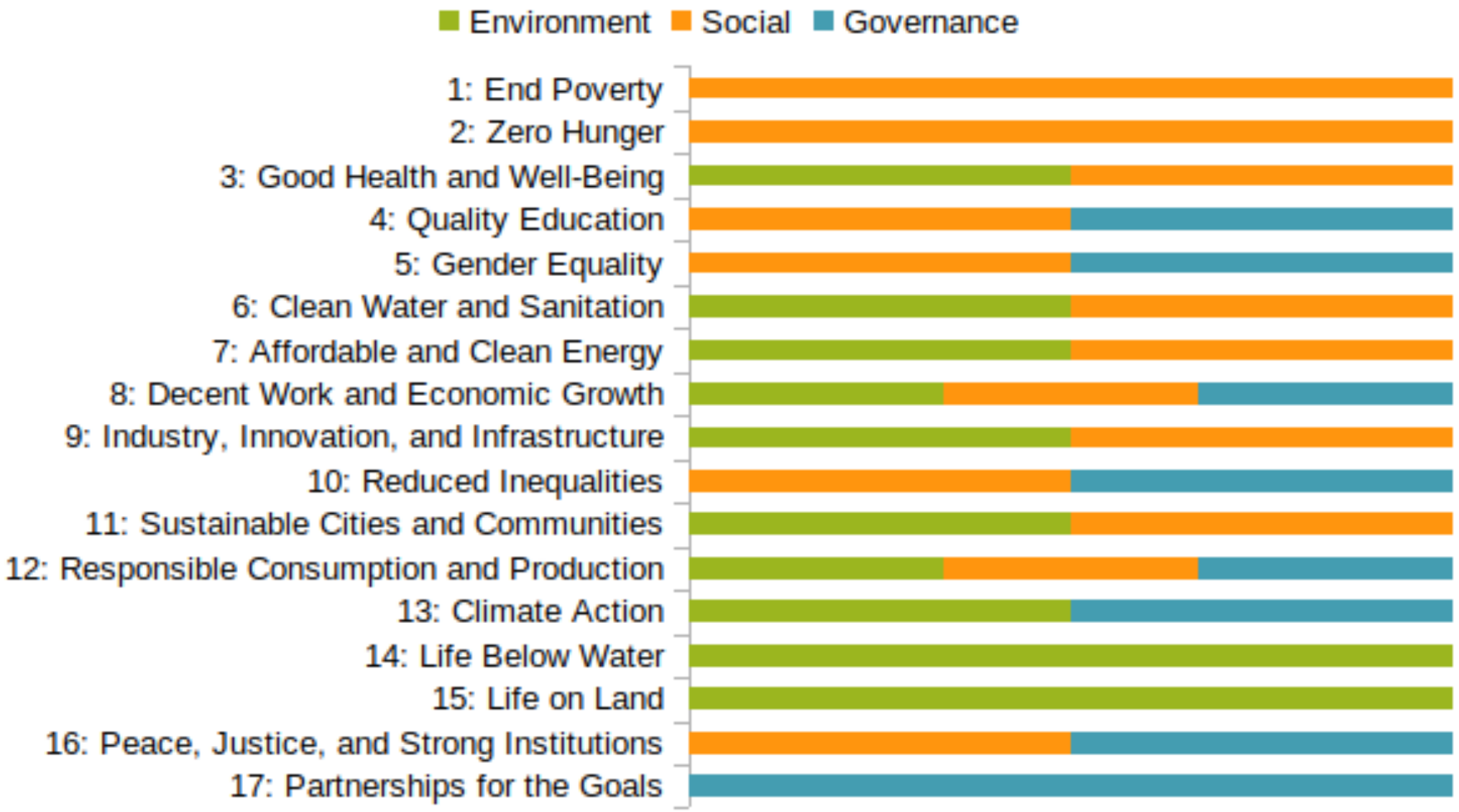
Planet

Source: Folke, Carl, Reinette Biggs, Albert V. Norström, Belinda Reyers, and Johan Rockström. "Social-ecological resilience and biosphere-based sustainability science."Ecology and Society 21, no. 3 (2016).

ESG to 17 SDGs

Source: Henrik Skaug Sætra (2021) "A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs." Sustainability 13, no. 15 (2021): 8503.

ESG to 17 SDGs



Source: https://sustainometric.com/esg-to-sdgs-connected-paths-to-a-sustainable-future/

AI for Social Good (AI4SG)
AI for Sustainable Development
AI4SG 10 Guidelines

- **AI Technology (G1, G2, G3)**
- **Applications (G4, G5, G6, G7, G8)**
- **Data Handling (G9, G10)**

Source: Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020) "AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.

AI4SG 10 Guidelines
AI Technology (G1, G2, G3)

- **G1: Expectations of what is possible with AI need to be well-grounded.**
- **G2: There is value in simple solutions.**
- **G3: Applications of AI need to be inclusive and accessible, and reviewed at every stage for ethics and human rights compliance.**

Source: Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020) "AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.

AI4SG 10 Guidelines
Applications (G4, G5, G6, G7, G8)

- **G4: Goals and use cases should be clear and well-defined.**
- **G5: Deep, long-term partnerships are required to solve large problems successfully.**
- **G6: Planning needs to align incentives, and factor in the limitations of both communities.**
- **G7: Establishing and maintaining trust is key to overcoming organisational barriers.**
- **G8: Options for reducing the development cost of AI solutions should be explored.**

Source: Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020) "AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.

AI4SG 10 Guidelines
Data Handling (G9, G10)

- **G9: Improving data readiness is key.**
- **G10: Data must be processed securely, with utmost respect for human rights and privacy.**

Source: Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020) "AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.

Papers with Code
State-of-the-Art (SOTA)

Computer Vision

Semantic Segmentation

185 benchmarks

3397 papers with code

Image Classification

390 benchmarks

2778 papers with code

Object Detection

269 benchmarks

2559 papers with code

Contrastive Learning

2 benchmarks

1119 papers with code

Image Generation

208 benchmarks

1097 papers with code

See all 1415 tasks

Natural Language Processing

Language Modelling

458 benchmarks

2248 papers with code

Question Answering

181 benchmarks

1818 papers with code

Machine Translation

78 benchmarks

1721 papers with code

Sentiment Analysis

87 benchmarks

1040 papers with code

Text Generation

242 benchmarks

931 papers with code

See all 664 tasks

https://paperswithcode.com/sota

Summary

- Philosophy, Ethics, and Safety of AI
  - The Limits of AI
  - Can Machines Really Think?
  - The Ethics of AI
- The Future of AI
  - AI Components
  - AI Architectures

References

- Stuart Russell and Peter Norvig (2020), Artificial Intelligence: A Modern Approach, 4th Edition, Pearson.
- Denis Rothman (2024), Transformers for Natural Language Processing and Computer Vision - Third Edition: Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3, 3rd ed. Edition, Packt Publishing
- Aurélien Géron (2022), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 3rd Edition, O'Reilly Media.
- Steven D'Ascoli (2022), Artificial Intelligence and Deep Learning with Python: Every Line of Code Explained For Readers New to AI and New to Python, Independently published.
- Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. (2025) "Yolov9: Learning what you want to learn using programmable gradient information." In European Conference on Computer Vision, pp. 1-21. Springer, Cham.
- Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou et al. (2025) "AI Alignment: A Contemporary Survey." ACM Computing Surveys (2025).
- Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. (2024) "Aligning cyber space with physical world: A comprehensive survey on embodied ai." arXiv preprint arXiv:2407.06886.
- Mihaela Constantinescu, and Muel Kaptein. (2025) "Responsibility gaps, LLMs & organisations: Many agents, many levels, and many interactions." Science and Engineering Ethics 31, no. 6 (2025): 36.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. (2021) "Learning transferable visual models from natural language supervision." In International Conference on Machine Learning, pp. 8748-8763. PMLR.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. (2021) "Vilt: Vision-and-language transformer without convolution or region supervision." In International Conference on Machine Learning, pp. 5583-5594. PMLR.
- Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. (2022) "Attention mechanisms in computer vision: A survey." Computational Visual Media ,:1-38.
- Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann.(2020) "Blazepose: On-device real-time body pose tracking." arXiv preprint arXiv:2006.10204.
- Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna et al.(2025) "Gemini robotics: Bringing ai into the physical world." arXiv preprint arXiv:2503.20020 (2025).
- Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay et al. (2025) "Cosmos world foundation model platform for physical ai." arXiv preprint arXiv:2501.03575 (2025).
- Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu et al. (2025) "Foundation models in robotics: Applications, challenges, and the future." The International Journal of Robotics Research 44, no. 5 (2025): 701-739.
- Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. (2025) "Vision-language-action models for robotics: A review towards real-world applications." IEEE Access (2025).
- Muhayy Ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. (2025) "Vision language action models in robotic manipulation: A systematic review." arXiv preprint arXiv:2507.10672 (2025)
- Rui Shao, Wei Li, Lingsen Zhang, Renshan Zhang, Zhiyang Liu, Ran Chen, and Liqiang Nie. (2025) "Large vlm-based vision-language-action models for robotic manipulation: A survey." arXiv preprint arXiv:2508.13073 (2025).
- Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020) "AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.
- Fernando Gonzalez, Zhijing Jin, Jad Beydoun, Bernhard Schölkopf, Tom Hope, Rada Mihalcea, and Mrinmaya Sachan (2022). "How Is NLP Addressing the 17 UN Sustainability Goals? A Challenge Set of Social Good Paper Classification and Information Extraction."
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. (2024) "A survey on multimodal large language models." National Science Review (2024): nwae403.
- Jiayi Kuang, Jingyou Xie, Haohao Luo, Ronghao Li, Zhe Xu, Xianfeng Cheng, Yinghui Li, Xika Lin, and Ying Shen. (2024) "Natural Language Understanding and Inference with MLLM in Visual Question Answering: A Survey." arXiv preprint arXiv:2411.17558.
- Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. (2023) "Trustworthy AI: From principles to practices." ACM Computing Surveys 55, no. 9 (2023): 1-46.
- Min-Yuh Day (2025), Python 101, https://tinyurl.com/aintpupython101