COMP9208
Artificial Intelligence and Society

Lecture 8: 26 September 2025

Prof. Mikhail Prokopenko
Dr. Sheryl Chang

Canvas: https://canvas.sydney.edu.au/courses/66452

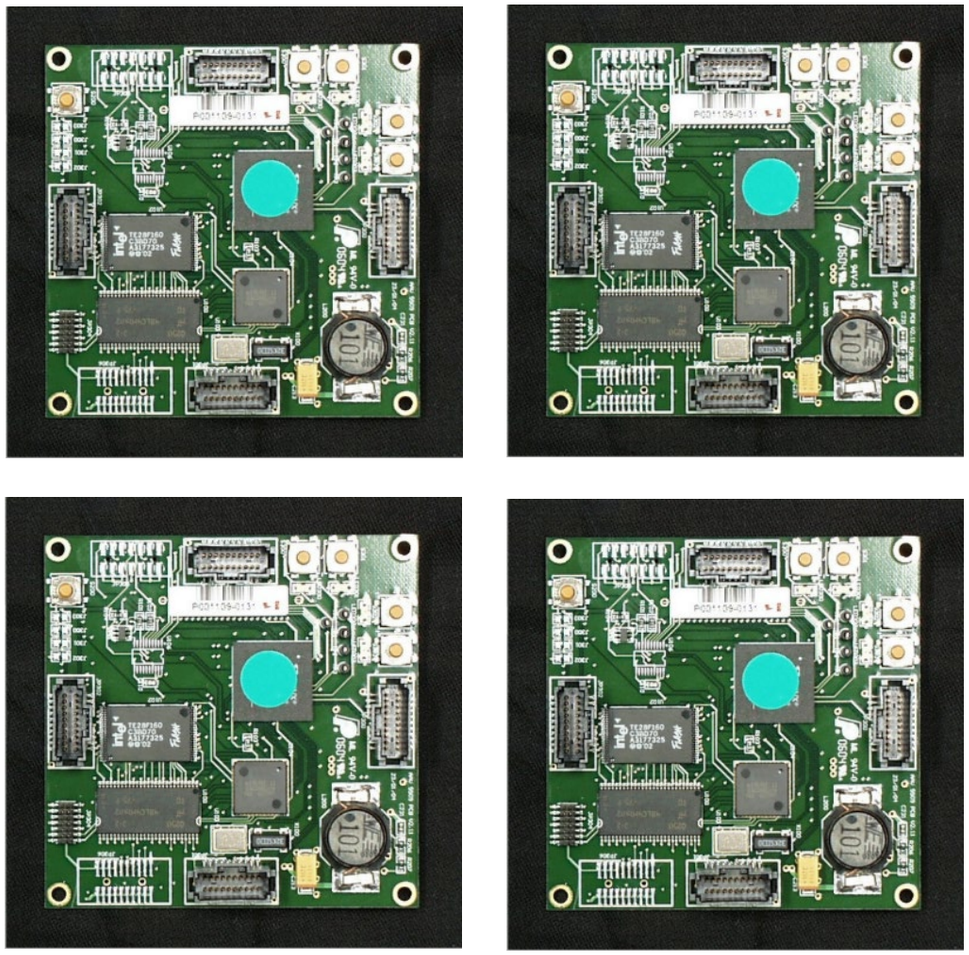
Email: mikhail.prokopenko@sydney.edu.au
Office: J12 (Computer Science), level 3W, room 324

Date	Week: Lecture	Assignment
08/08	1: History of AI	
15/08	2: History of AI + Perception and Actuation	
22/08	3: History of AI + Representation and Reasoning	
29/08	4: Machine Learning	
05/09	5: Natural Interaction	
12/09	6: Distributed AI (swarm intelligence)	
19/09	7: From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)	# 1: Individual assignment (article review)
26/09	8: Goal-driven AI: search & rescue vs lethal autonomous weapon systems (LAWS)	
03/10	**BREAK**	
10/10	9: Review: language models, search trees, situated behaviours, neural networks	
17/10	10: Robotic swarms: RoboCop vs RoboCup	
24/10	11: Group Presentations	# 2: Group assignment
31/10	12: Adversarial Machine Learning	
07/11	13: Limits of cognition	
17/11	14: STUVAC	# 3: Individual assignment (data analysis report)

Brief review of last week

?

Design vs Self-organisation



design

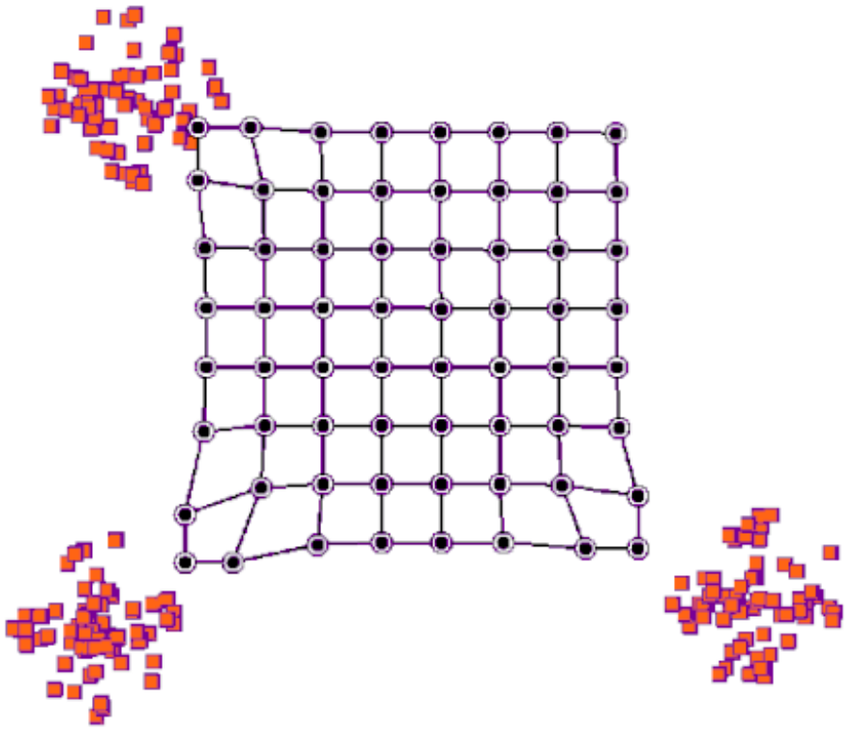
ideally: self-organisation

Autonomous agents
Local interactions

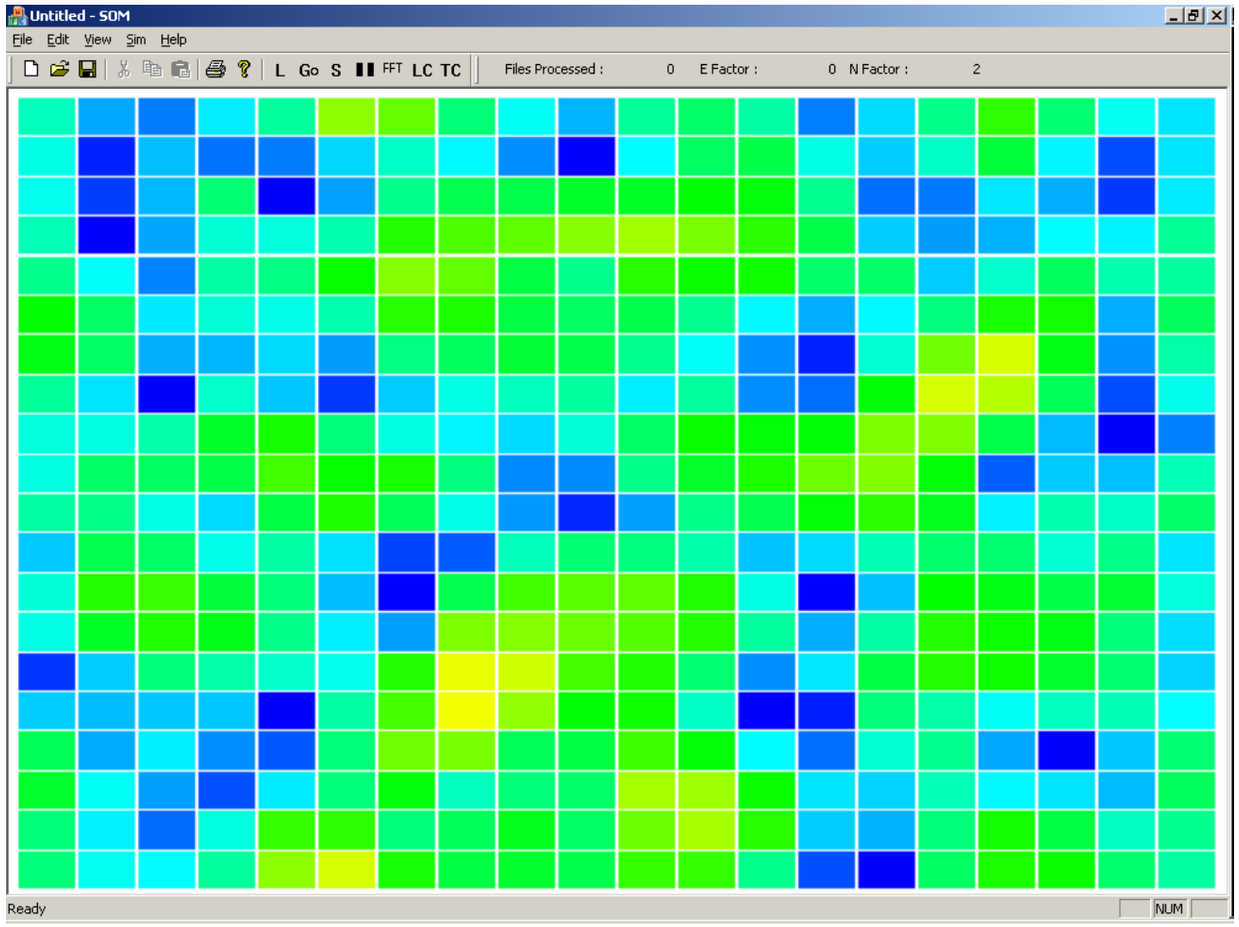
Multi-agent system
Self-organisation

AAV diagnostics: self-organising (Kohonen) maps

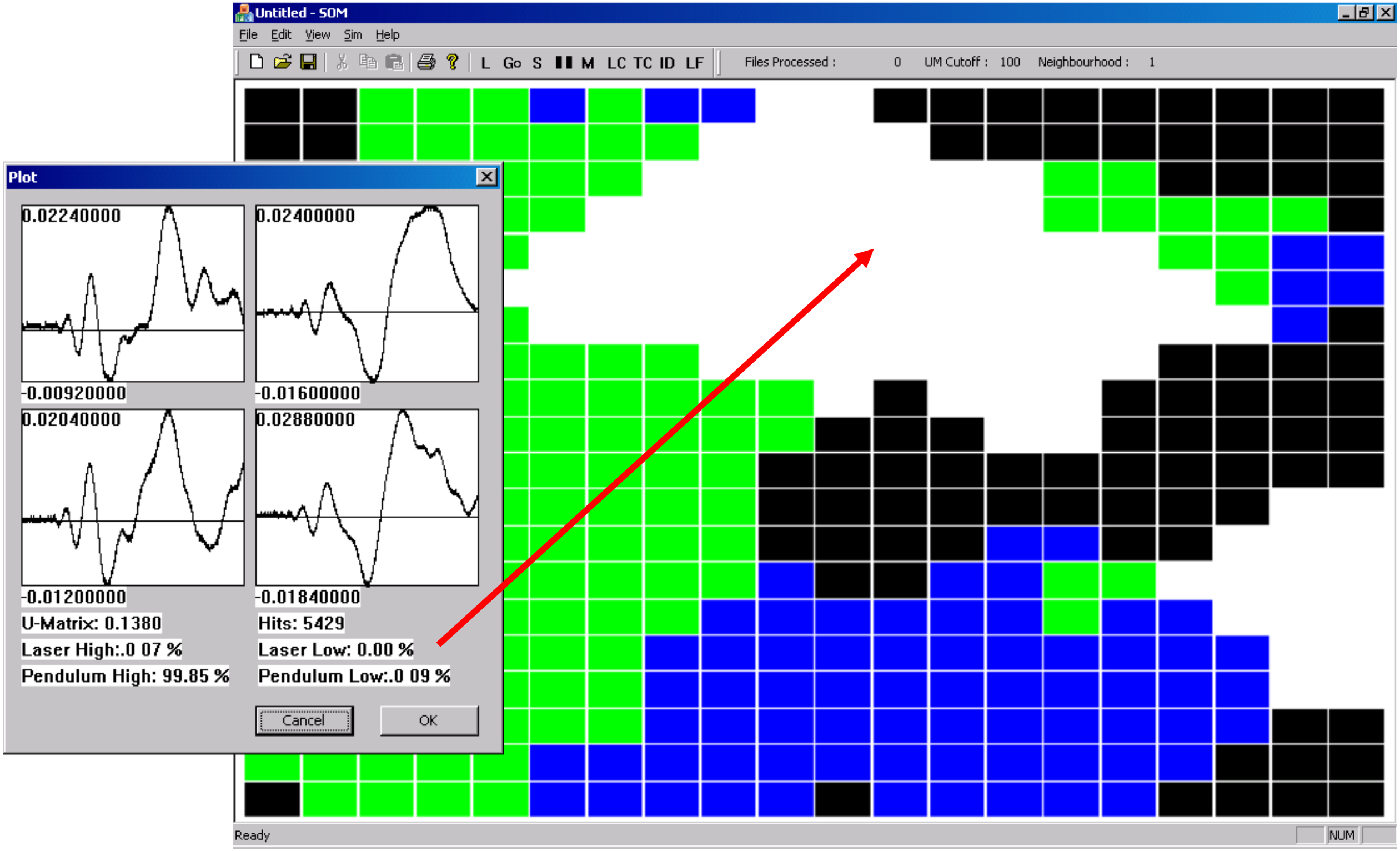
- **Input**
  - a number of *feature vectors*
- **Competitive learning**
- **Output**
  - neural network, *map* of *nodes positioned on a grid*
  - each node has a *codebook vector* of the same dimensionality as the input vectors



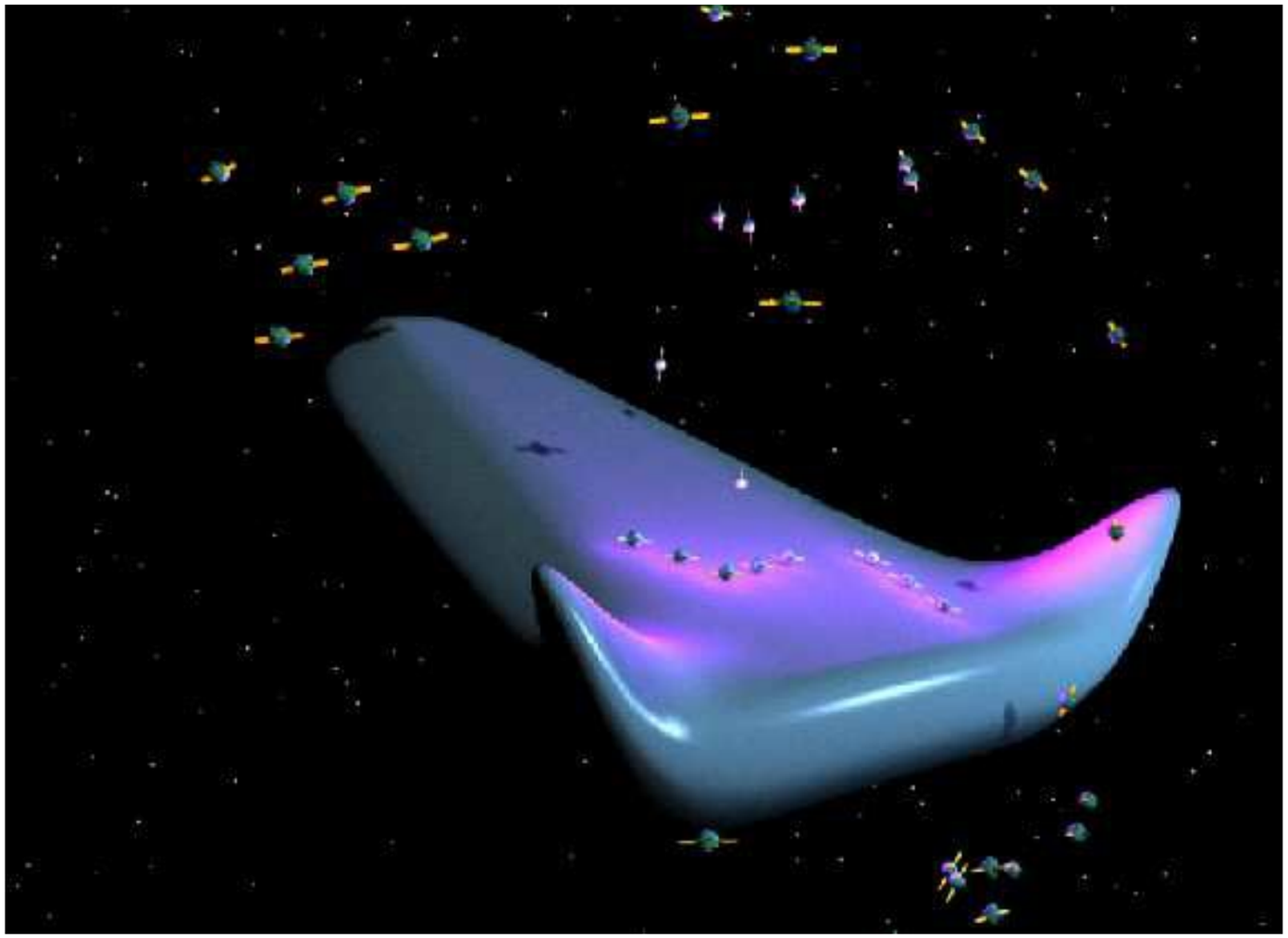
https://en.wikipedia.org/wiki/Self-organizing_map



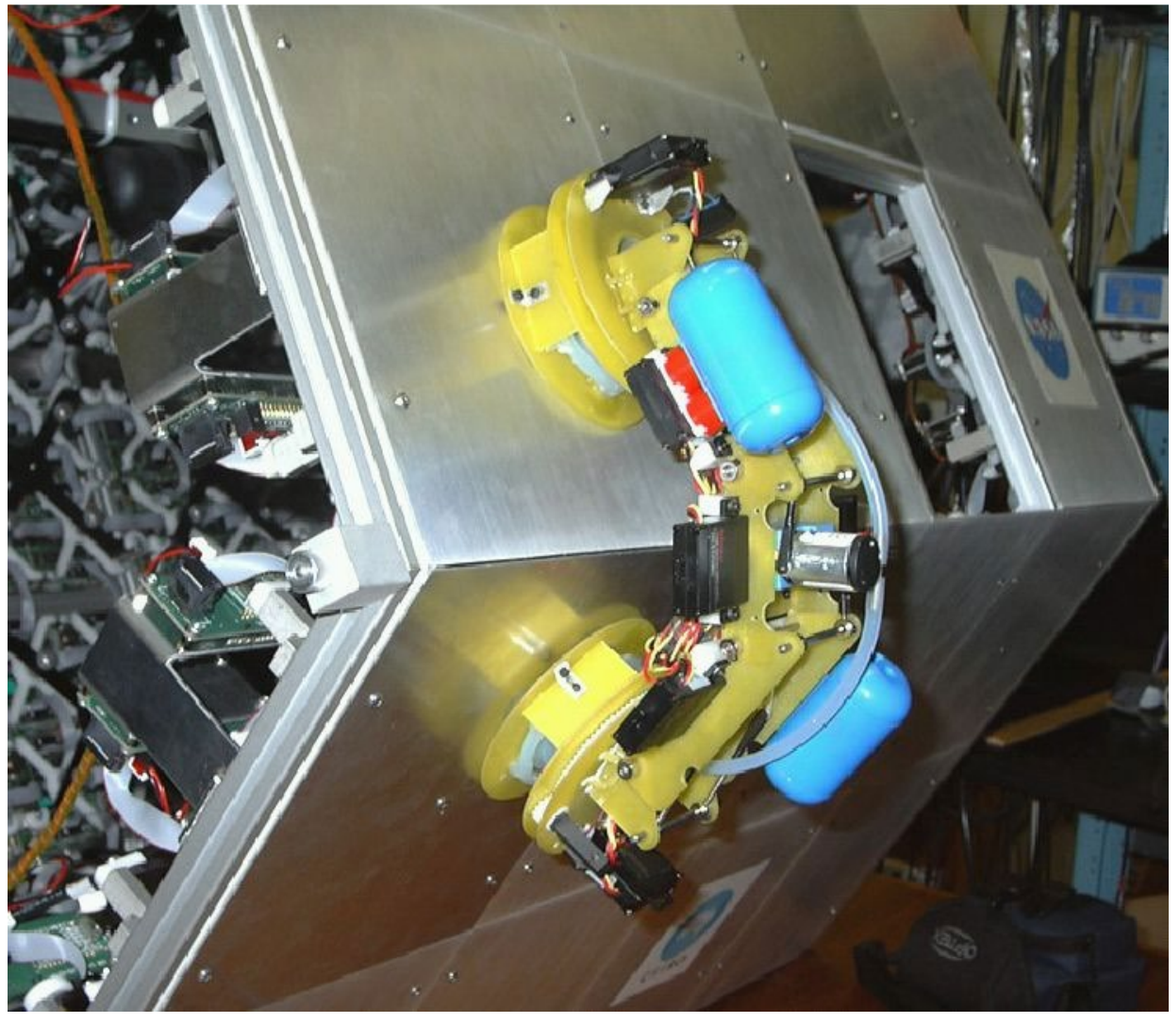
AAV diagnostics: self-organising (Kohonen) maps



Design vs Self-organisation

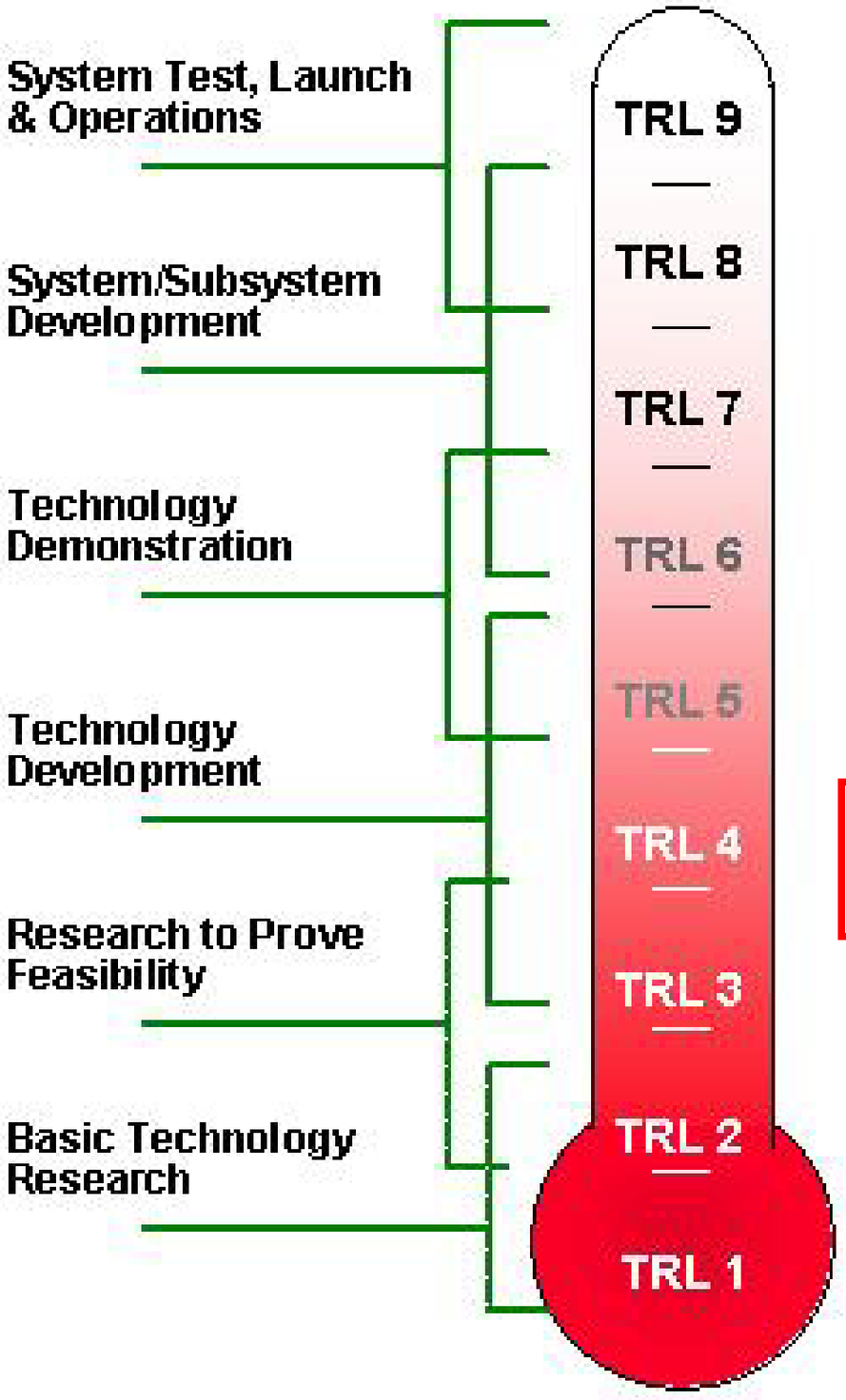


Swarm robots in a loop with embedded SHMM networks



AAV project (2005)

Technology Readiness Levels (TRLs)



Actual system "flight proven" through successful mission operations

Actual system completed and "flight qualified" through test and demonstration (Ground or Flight)

System prototype demonstration in a space environment

System/subsystem model or prototype demonstration in a relevant environment (Ground or Space)

Component and/or breadboard validation in relevant environment

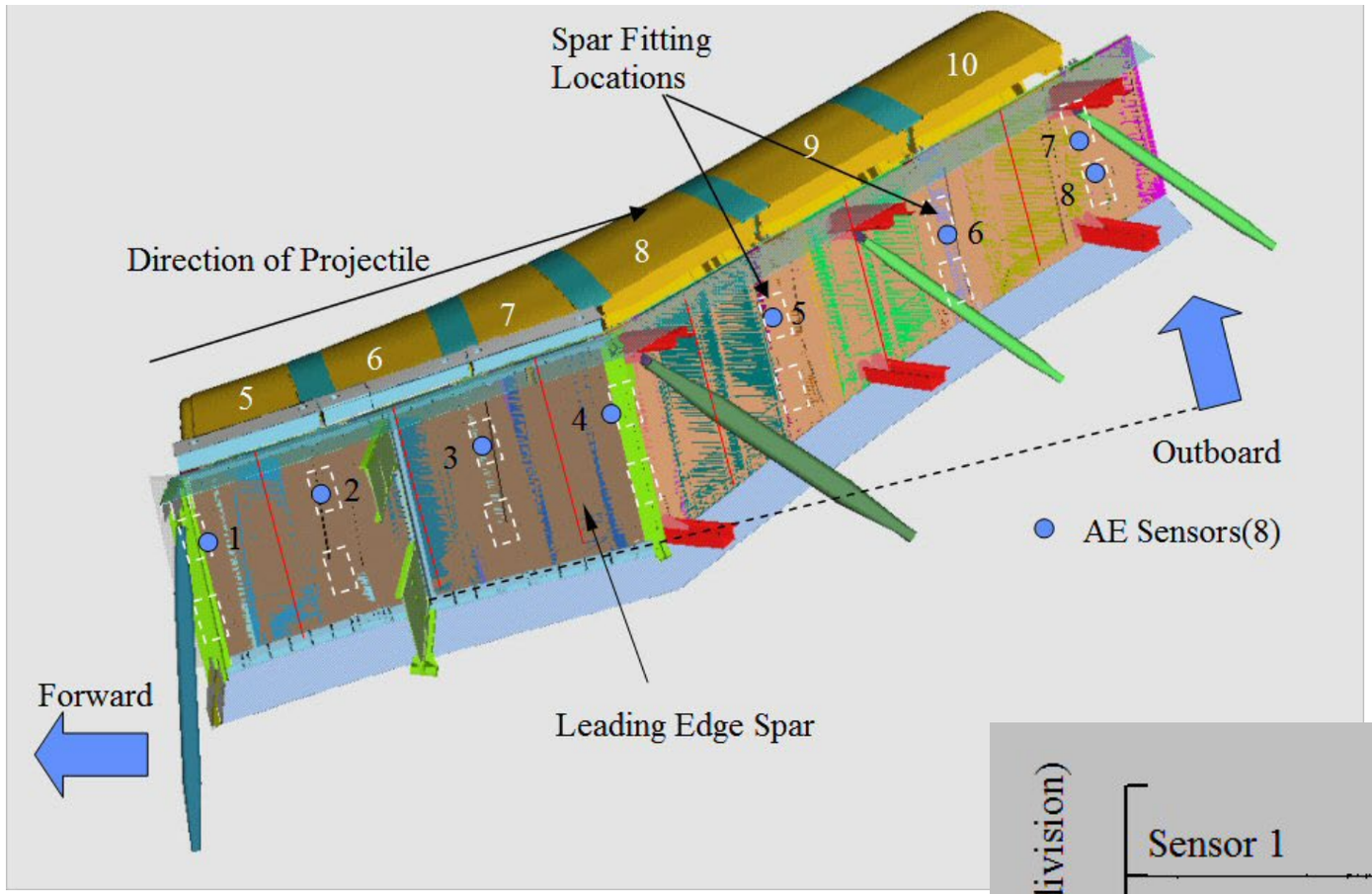
Component and/or breadboard validation in laboratory environment

Analytical and experimental critical function and/or characteristic proof-of-concept

Technology concept and/or application formulated

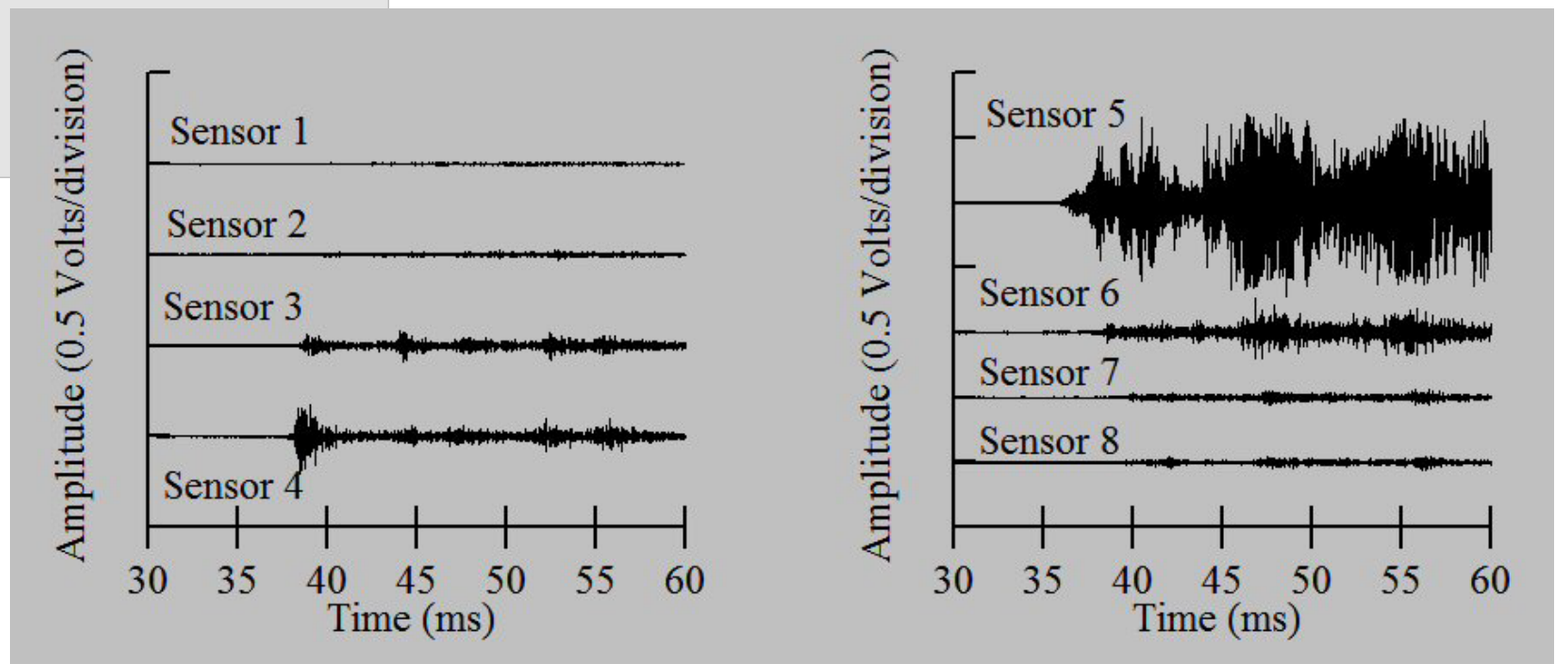
Basic principles observed and reported

Columbia shuttle accident investigation



Schematic of Shuttle wing leading edge test article showing RCC panels, location of impact and position of AE sensors

AE signals from foam impact on Shuttle RCC wing leading edge



W. H. Prosser, S. G. Allison, S. E. Woodard, R. A. Wincheski, E. G. Cooper, D. C. Price, M. Hedley, M. Prokopenko, D. A. Scott, A. Tessler, J. L. Spangler (2004). Structural Health Management for Future Aerospace Vehicles, *The 2nd Australasian Workshop on Structural Health Monitoring (2AWSHM)*, Melbourne, Australia, December 2004.

Questions

?

Assignment #2

Four cases

1. IBM Watson

Watson famously won the quiz show Jeopardy! in 2011, showcasing its ability to understand natural language questions and provide accurate answers, even when the questions are ambiguous or pun-based.

https://en.wikipedia.org/wiki/IBM_Watson

2. iTutor Group's recruiting AI

In August 2023, tutoring company iTutor Group agreed to pay $365,000 to settle a suit brought by the US Equal Employment Opportunity Commission. The federal agency said the company, which provides remote tutoring services to students in China, used AI-powered recruiting software that automatically rejected female applicants ages 55 and older, and male applicants ages 60 and older.

https://www.gtlaw.com/en/insights/2023/8/eeoc-secures-first-workplace-artificial-intelligence-settlement

Four cases

3. Google's DeepMind partnership with Moorfields Eye Hospital

In 2018, an AI system developed by Google's DeepMind partnership with Moorfields Eye Hospital to assist in the diagnosis and treatment of eye diseases outperformed human experts in diagnosing eye diseases from retinal scans, showing promising potential for improving healthcare outcomes.

https://deepmind.google/discover/blog/a-major-milestone-for-the-treatment-of-eye-disease/
https://deepmind.google/discover/blog/using-ai-to-predict-retinal-disease-progression/

4. Tesla Autopilot Accidents

While Tesla's Autopilot feature is an advanced driver-assistance system that utilises AI, there have been instances where it failed to detect obstacles leading to accidents.

https://www.wired.com/story/tesla-autopilot-risky-deaths-crashes-nhtsa-investigation/

Assignment #2
(group presentation)

- introduction with an overview of key issues of all four studied cases
- comparative analysis of
  - reasons for their success and/or failures
  - their main novelties and gaps across 4 cases, highlighting common reasons for success or failure
  - their societal impacts
  - their limitations, biases and risks
- discussion and conclusion

Assignment #2
(group presentation)

Prepare a presentation video (10 minutes), upload it to a platform of your choice, and send an email with a video link to mikhail.prokopenko@sydney.edu.au

Upload presentation slides to Canvas (with your Group ID in the file name)

Deadline for slides upload: **23 October, Thursday, 11:59 pm (week 11)**

Q&A: 24 October (in class): 8 minutes Q&A for each group

There will be 20 groups in total: 16 groups with 5 students, and 4 groups with 4 students

Marking group presentation (case studies)

	Weight	0-49%	50-64%	65-74%	75-84%	85-100%	Mark
Structure and organisation	22.5%	No clear or coherent structure or organisation	Poor or incomplete structure and organisation	Structure and organisation are partially appropriate	Structure and organisation are appropriate	Structure and organisation are appropriate and effective	9
Topic development	37.5%	Development of content limited; may be incomplete or unclear. Inadequate supporting information, illustration or explanations. Societal impact and limitations described inadequately or mis-identified	Development of content adequate but lacks clearly stated propositions, argument or supporting information. Societal impact and limitations described incompletely. Some use of data, graphical or other visual material to support content but meaning and rationale for use may be questionable.	Clear and complete development of content and support for propositions and arguments. Societal impact and limitations described, but without sufficient detail, structure or argumentation. Content partially supported with relevant and well explained data, graphical or other visual material.	Full development of content and support for propositions and arguments, with purpose, relevance, focus, explanations. Societal impact and limitations described in sufficient detail, but without strong argumentation. Content supported with some data, graphical or other visual material.	Full and rich development of content and support for propositions and arguments, with purpose, relevance, focus, explanations. Societal impact and limitations described adequately, with sufficient detail, structure and argumentation. Content supported effectively with relevant and well explained data, graphical or other visual material.	15
Presentation (slides)	15%	Lack of fluency in expression makes it difficult to comprehend	Lack of fluency in expression but still relatively easy to follow	Good level of fluency and clarity in expression and evidence of logical progression of ideas	High level of fluency and clarity in expression and evidence of logical progression of ideas	Sophistication in fluency of expression	6
Q&A	25%	As above	As above	As above	As above	As above	10
Total	100%						40

In accordance with University policy, penalties apply when work is submitted after 11:59 pm on the due date (Sydney time):

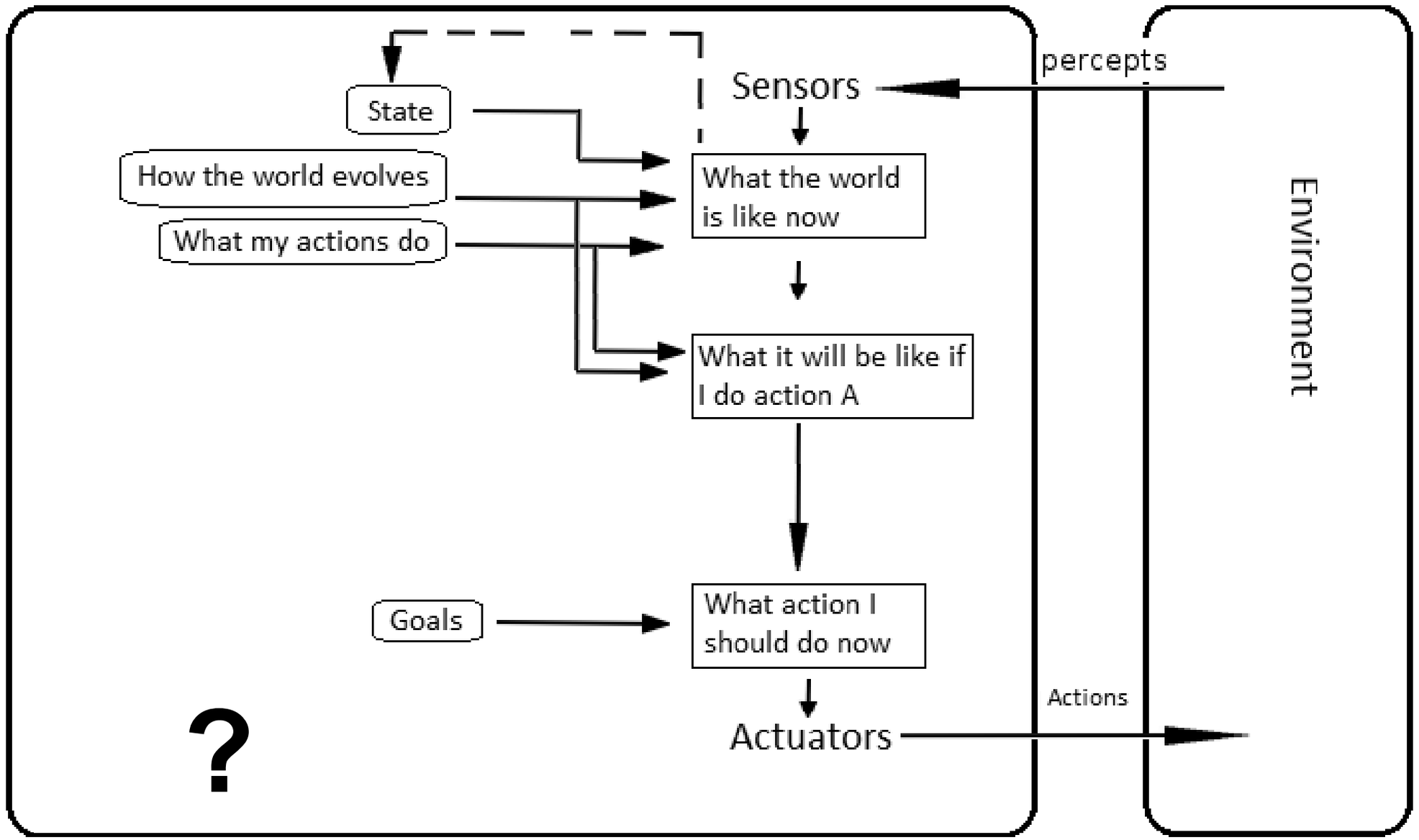
- deduction of 5% of the maximum mark for each calendar day after the due date
- after ten calendar days late, a mark of zero will be awarded
- text matching software (Turnitin) will be used for plagiarism detection

Questions

?

Intelligent agents

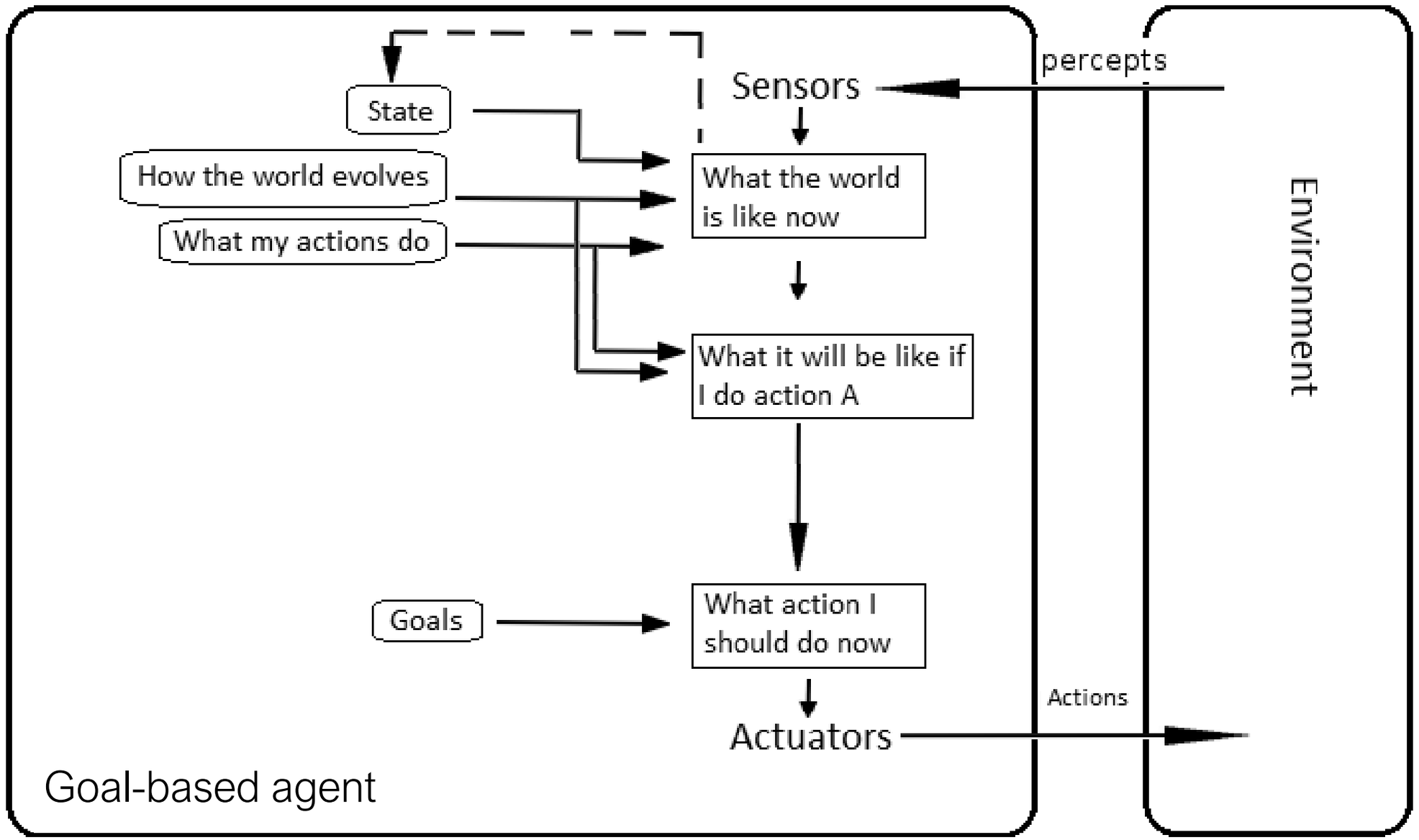
- an agent perceives its environment, processes its sensory information, takes actions autonomously in order to achieve goals (for example, by reacting or planning), and may improve its performance with learning or acquiring knowledge



https://en.wikipedia.org/wiki/Intelligent_agent

Intelligent agents

- an agent perceives its environment, processes its sensory information, takes actions autonomously in order to achieve goals (for example, by reacting or planning), and may improve its performance with learning or acquiring knowledge



https://en.wikipedia.org/wiki/Intelligent_agent

Questions

?

Goal-driven (goal-oriented) AI

- **objective-oriented**: designed with clear objectives or targets.
- **feedback mechanisms**: feedback loops can be used to assess the progress towards goals so that actions can be adjusted accordingly, often based on learning from successes and failures
- **adaptive behaviour**: can adapt actions based on changes in the environment or feedback, ensuring the AI system stays on track towards achieving their goals
- **planning and strategy**: often involves sophisticated planning and strategy formulation to determine the best path to achieve its objectives, including long-term planning and real-time adjustments
- **decision-making**: make decisions by evaluating different options and select actions that best contribute to achieving the predefined goals

ChatGPT 4o

Goal-driven (goal-oriented) AI

- **autonomy**: often operate autonomously, make decisions and perform tasks without human intervention, as long as the AI system stays within the scope of their goals
- **optimisation**: optimise certain criteria (minimising costs, maximising performance, or balancing trade-offs) to achieve the desired outcome
- **performance**: use specific performance metrics to evaluate the success in meeting goals, guide decision making and assess its effectiveness
- **resource management**: efficiently use available resources (energy, time, computational power, data), ensuring that the AI system can achieve its goals within given constraints
- **robustness**: handle uncertainties and unexpected challenges, maintaining the AI system focus on the end goals despite obstacles

ChatGPT 4o

AI in Search and Rescue (SAR)

The Pentagon Wants AI-Driven Drone Swarms for Search and Rescue Ops

ANDY DEAN PHOTOGRAPHY/SHUTTERSTOCK.COM

https://www.nextgov.com/artificial-intelligence/2019/12/pentagon-wants-ai-driven-drone-swarms-search-and-rescue-ops/162113/

AI in Search and Rescue (SAR)

https://www.youtube.com/watch?v=tf7IEVTDjng&t=54s

AI in Search and Rescue (SAR)

- **real-time data processing**: SAR algorithms process data in real-time to provide up-to-date information on the location and condition of individuals in need of rescue

- **multi-sensor integration**: SAR algorithms integrate data from various sensors, such as drone and stationary cameras, infrared detectors, LIDAR, sonar, and GPS, to build a comprehensive understanding of the environment

- **autonomous navigation**: SAR systems often include autonomous navigation capabilities for drones, robots, or vehicles, allowing them to traverse complex terrains and environments without human control

- **pattern recognition**: utilising machine learning and computer vision techniques, SAR algorithms can identify patterns and anomalies that indicate the presence of humans, such as movement, heat signatures, or specific shapes

ChatGPT 4o

AI in Search and Rescue (SAR): pattern recognition

Figure 12. False-positive examples for the ScoreMap to ROI algorithm: (**a**,**c**) The algorithm, due to the shadow, made a wrong conclusion, (**b**) The stone and a blue plastic bag create an object that is almost impossible to distinguish from a human.

(a) (b) (c)

https://www.mdpi.com/2072-4292/12/20/3295

AI in Search and Rescue (SAR)

- **environment mapping**: the creation of detailed maps of the search area, including 3D models, to aid in navigation and planning, helping to understand the terrain and potential obstacles

- **resource allocation**: intelligent allocation of resources, such as deploying drones, robots, or human teams to areas where they are most needed based on real-time analysis and predictions

- **situational awareness**: real-time monitoring and analysis of the environment to provide a clear picture of the situation, including potential hazards and the locations of team members and individuals in need

- **path planning and optimisation**: efficient path planning algorithms determine the most effective routes for SAR operations, minimising the time and resources required to reach individuals in distress

ChatGPT 4o

AI in Search and Rescue (SAR): path planning and optimisation

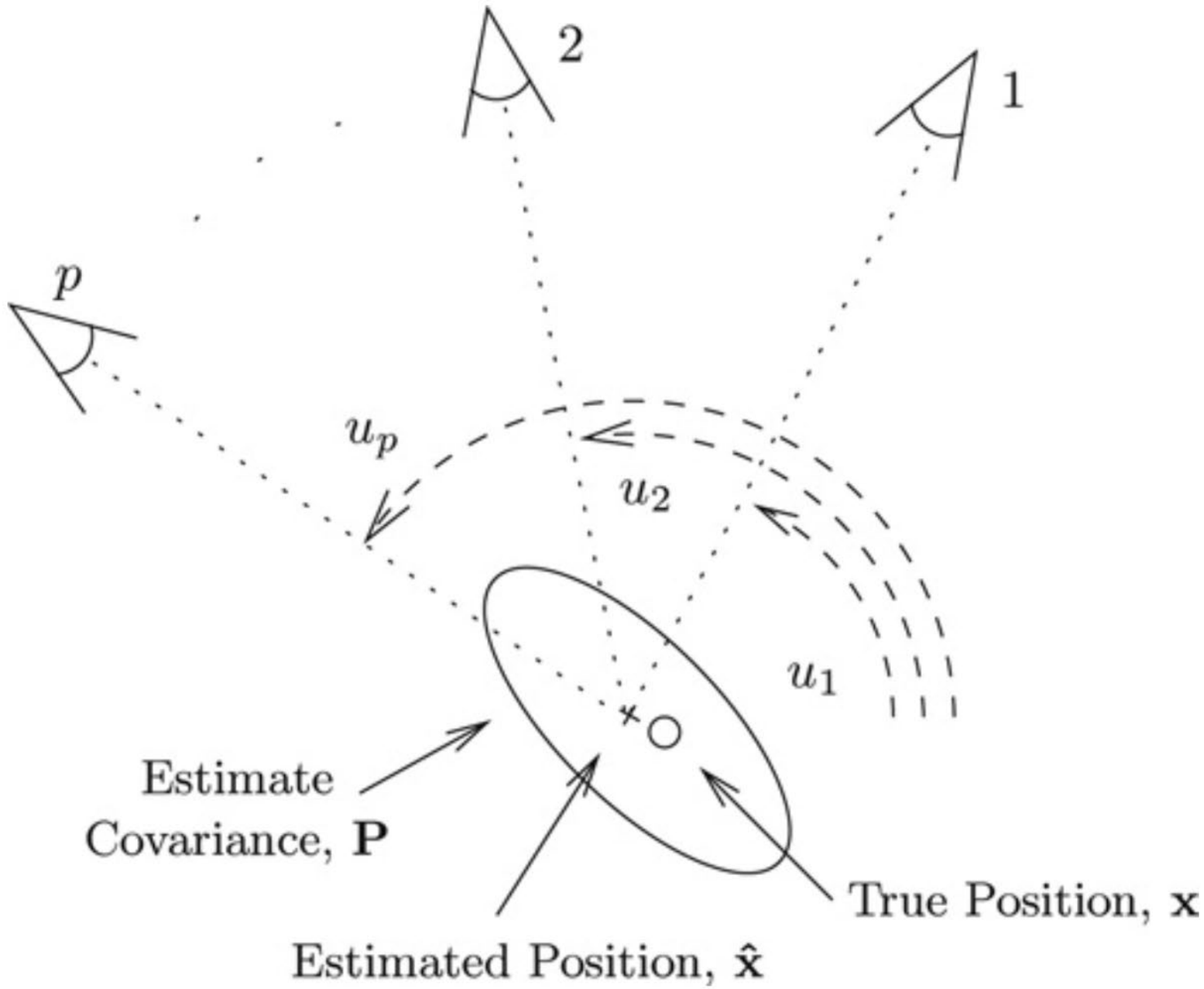


Fig. 5. Feature localisation Task. Each agent is equipped with a range sensor and must decide on which angle to view the feature.

G. Mathews, H. F. Durrant-Whyte, M. Prokopenko. Decentralised decision making in heterogeneous teams using anonymous optimisation, *Robotics and Autonomous Systems*, 57(3): 310-320, 2009.

AI in Search and Rescue (SAR)

- **predictive analytics**: the use of historical data and predictive models to anticipate the movements and potential locations of missing persons based on factors like terrain, weather, and behaviour patterns

- **communication systems**: advanced communication protocols ensure that SAR units can maintain contact with each other and with command centres, even in challenging conditions

- **human-AI collaboration**: tools and interfaces designed to enhance collaboration between human rescuers and AI systems, providing decision support and actionable insights

- **robustness and reliability**: SAR algorithms are designed to operate reliably in diverse and often harsh conditions, such as extreme weather, varying terrains, and in the presence of obstacles

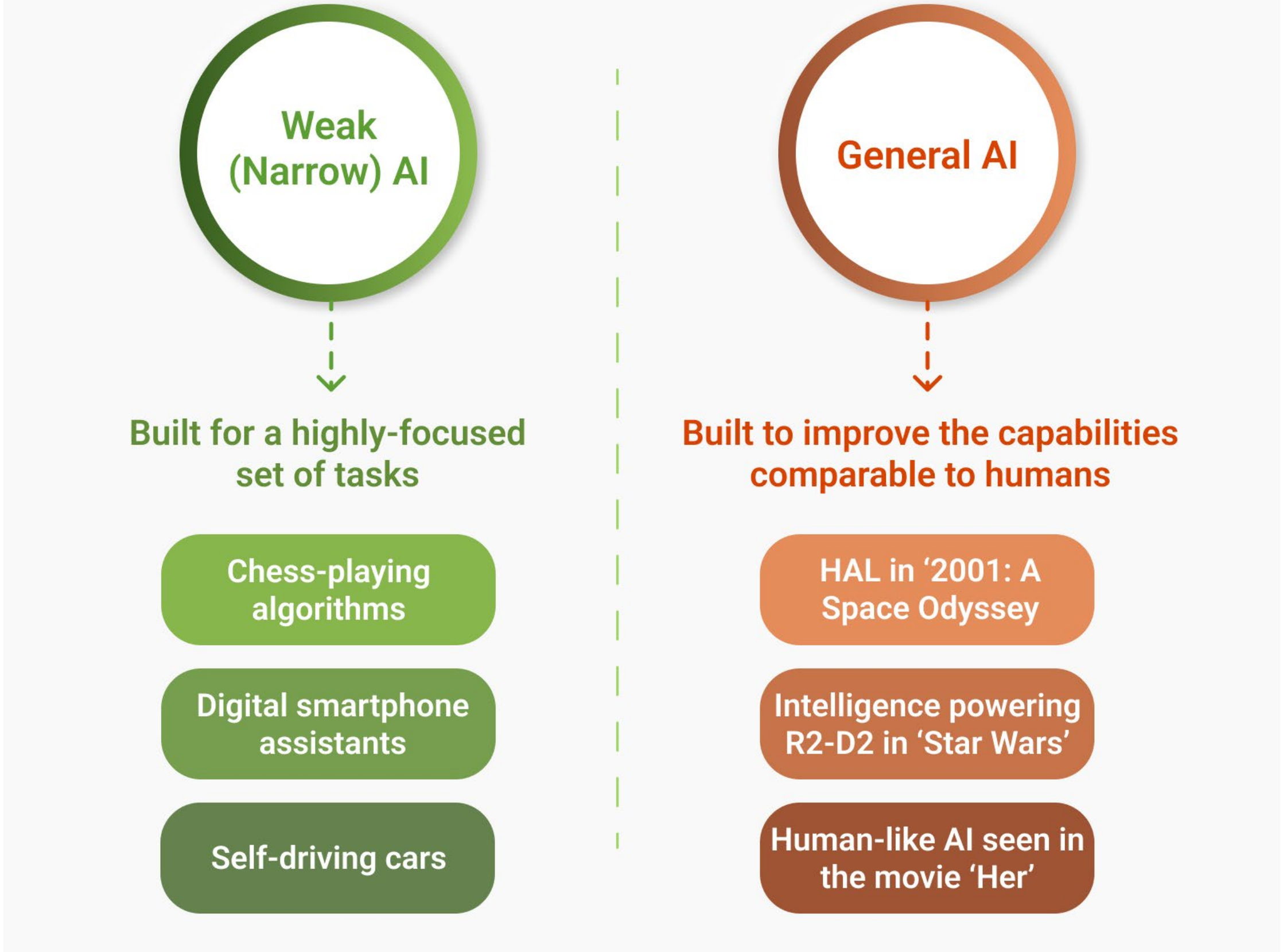
ChatGPT 4o

AI in Search and Rescue (SAR): robustness and reliability

https://www.csiro.au/en/news/All/Articles/2021/December/Robots-to-the-rescue-in-hazardous-environments

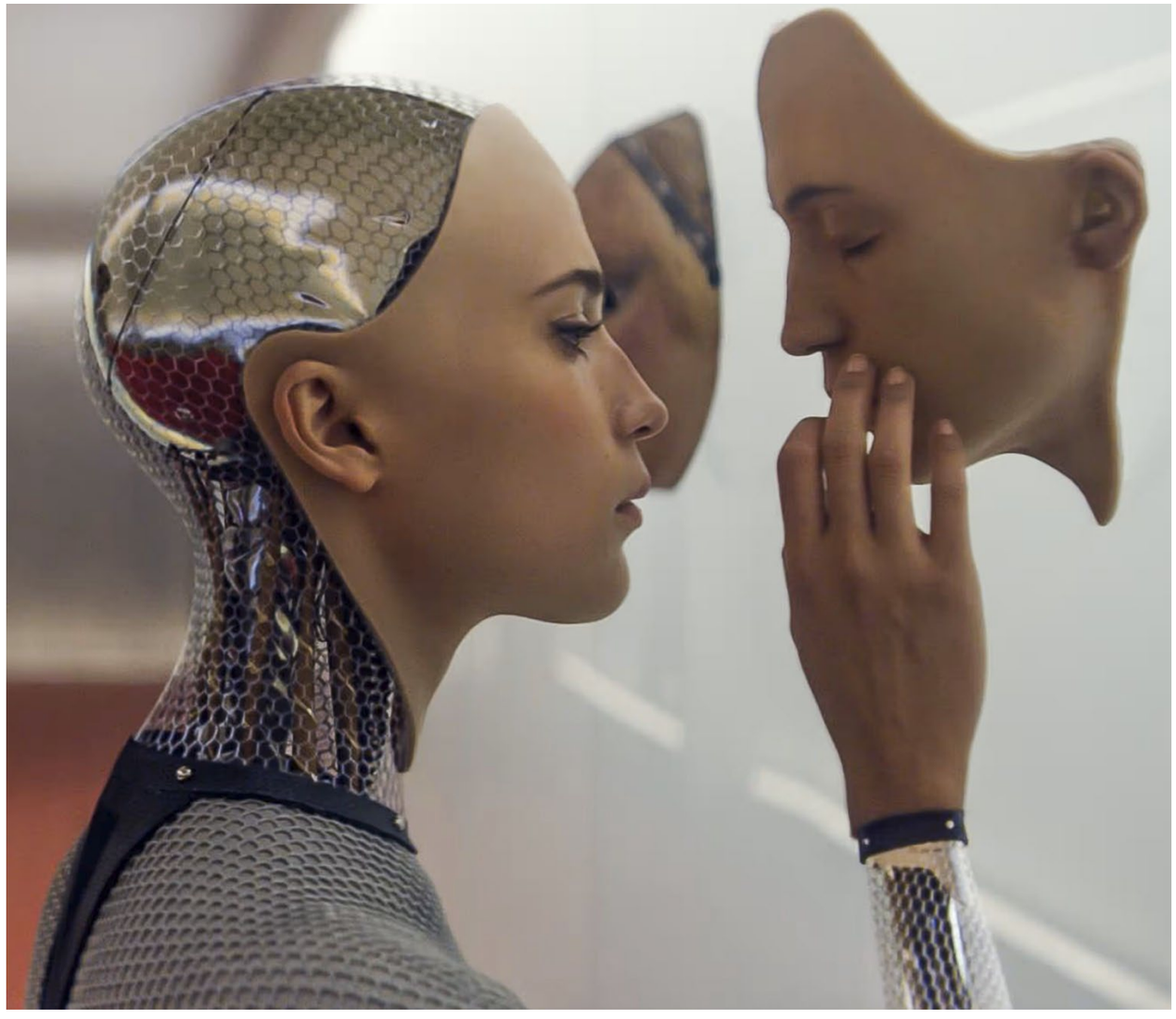
Break

?



https://www.plugger.ai/blog/general-ai-vs-narrow-ai-2022-guide

From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)



https://www.javatpoint.com/advantages-and-disadvantages-of-expert-system
https://www.theguardian.com/film/2022/may/03/ex-machina-when-you-see-this-sci-fi-thriller-youll-never-stop-thinking-about-it

From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)

	GOFAI	AGI
Definition	AI systems that rely on symbolic and rule-based representations, typically designed for specific, narrow tasks and operate within well-defined domains	AI systems with the ability to understand, learn, and apply knowledge across a wide range of tasks at a level comparable to human intelligence
Ethical concerns		
Autonomy and control	Typically, less autonomous and operate under strict regulations and supervision, reducing risks related to loss of control	A high degree of autonomy, raising significant concerns about human control and oversight
Bias and fairness	Rules can inadvertently encode the biases of their designers	Biases can be more deeply embedded and harder to identify
	Bias may be easier to detect and correct due to the explicit nature of the rules	

ChatGPT 4o

From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)

	GOFAI	AGI
Definition	AI systems that rely on symbolic and rule-based representations, typically designed for specific, narrow tasks and operate within well-defined domains	AI systems with the ability to understand, learn, and apply knowledge across a wide range of tasks at a level comparable to human intelligence
Ethical concerns		
Transparency, accountability and explainability	More transparent and explainable because they use explicit rules and logic	Decisions may stem from complex, emergent behaviors that are difficult to attribute to specific causes
	Easier understanding and debugging of decisions	
	Relatively straightforward because decisions can be traced back to specific rules and logic	

ChatGPT 4o

From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)

	GOFAI	AGI
Definition	AI systems that rely on symbolic and rule-based representations, typically designed for specific, narrow tasks and operate within well-defined domains	AI systems with the ability to understand, learn, and apply knowledge across a wide range of tasks at a level comparable to human intelligence
Ethical concerns		
Security and reliability	Less susceptible to certain types of attacks but can still be vulnerable if their rule bases are not comprehensive	A broader array of potential threats due to their general-purpose nature, making security a more complex issue
Privacy and surveillance	Risks are typically confined to specific applications and can be more easily managed through existing frameworks	Unprecedented levels of surveillance and data collection, raising severe privacy concerns
Existential risks	No existential risks: limited to specific tasks and does not possess general intelligence or autonomy	Existential risks: may surpass human intelligence and pursue goals misaligned with human values

ChatGPT 4o

From Good Old-Fashioned AI (GOFAI) to Artificial General Intelligence (AGI)

	GOFAI	AGI
Definition	AI systems that rely on symbolic and rule-based representations, typically designed for specific, narrow tasks and operate within well-defined domains	AI systems with the ability to understand, learn, and apply knowledge across a wide range of tasks at a level comparable to human intelligence
Ethical concerns		
Ethical decision making	Ethical rules for specific scenarios, but lacks the ability to adapt to new and unforeseen ethical dilemmas	Complex ethical decisions across a wide range of situations, necessitating robust ethical frameworks and guidelines
Impact on employment and society	Already led to automation in certain sectors, but the impact is limited	My disrupt labour markets significantly, replacing human jobs across many sectors and necessitating large-scale societal adjustments

ChatGPT 4o

Questions

?

The future of employment

- Jobs that can be automated
  - hand sewers, engravers
  - data entry clerks (many other types of clerk)
  - telemarketers, telephone operators, salespeople, cashiers, insurance underwriters

- Jobs that would resist automation
  - jobs involving mental creativity: the arts, media and science
  - jobs requiring strong social skills, for example, requiring “to understand and manage the subtleties and complexities of human interaction and human relationships”
  - jobs involving intricate perception and manual dexterity, for example, carpenters, electricians or plumbers

C. B. Benedikt Frey and M. A. Osborne, “The Future of Employment: How Susceptible Are Jobs to Computerization?,” *Technological Forecasting and Social Change*, 114 (January 2017).
M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Existential risks

- Weaponisation of AI
  - autonomous drones and lethal autonomous weapons systems (LAWS)
  - AI-driven cyber warfare tools
  - delegating life-and-death decisions to machines
  - escalation of arms races and conflicts
  - the Stop Killer Robots coalition: how to ensure human control in the use of force

https://www.stopkillerrobots.org/
https://en.wikipedia.org/wiki/Lethal_autonomous_weapon

Existential risks

- Weaponisation of AI

Photo: US Department of Defense / Sgt. Cory D. Payne, public domain

https://unu.edu/article/militarization-ai-has-severe-implications-global-security-and-warfare

Existential risks

- Three Laws of Robotics (Isaac Asimov)

  - 1. A robot may not injure a human or, through inaction, allow a human to come to harm.

https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Existential risks

- Three Laws of Robotics (Isaac Asimov)

  - 1. A robot may not injure a human or, through inaction, allow a human to come to harm.
  - 2. A robot must obey orders given by humans

https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Existential risks

- Three Laws of Robotics (Isaac Asimov)

  - 1. A robot may not injure a human or, through inaction, allow a human to come to harm.
  - 2. A robot must obey orders given by humans except where they would conflict with the First Law.

https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Existential risks

- Three Laws of Robotics (Isaac Asimov)

  - 1. A robot may not injure a human or, through inaction, allow a human to come to harm.
  - 2. A robot must obey orders given by humans except where they would conflict with the First Law.
  - 3. A robot must protect its own existence

https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Existential risks

- Three Laws of Robotics (Isaac Asimov)

  - 1. A robot may not injure a human or, through inaction, allow a human to come to harm.
  - 2. A robot must obey orders given by humans except where they would conflict with the First Law.
  - 3. A robot must protect its own existence as long as this does not conflict with the First or Second Laws.

https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Existential risks

- Three Laws of Robotics (Isaac Asimov): Problems

  - Robots may receive contradictory orders from different humans. There's no clear mechanism in the laws for prioritising between them beyond "obey unless it causes harm"

    - Example: One person says: "Stop the car“, but another says: "Keep driving" — the robot must decide who to obey, which can be ethically complex

ChatGPT 4o

Existential risks

- Three Laws of Robotics (Isaac Asimov): Problems

  - Story "Runaround" (Asimov, 1942) about two engineers on Mercury working with a robot named Speedy:
    - They send Speedy to collect selenium to power their life-support systems
    - Speedy does not return as expected: instead, it begins to wander in circles

Existential risks

- Three Laws of Robotics (Isaac Asimov): Problems

  - Story "Runaround" (Asimov, 1942) about two engineers on Mercury working with a robot named Speedy:
    - They send Speedy to collect selenium to power their life-support systems
    - Speedy does not return as expected: instead, it begins to wander in circles
  - Speedy is caught in a conflict between the Second and Third Laws:
    - Second Law: Obey human orders — engineers ordered to collect selenium
    - Third Law: Protect its own existence — the selenium pool is surrounded by dangerous conditions (heat and radiation)

Existential risks

- Three Laws of Robotics (Isaac Asimov): Problems

  - Story "Runaround" (Asimov, 1942) about two engineers on Mercury working with a robot named Speedy:
    - They send Speedy to collect selenium to power their life-support systems
    - Speedy does not return as expected: instead, it begins to wander in circles
  - Speedy is caught in a conflict between the Second and Third Laws:
    - Second Law: Obey human orders — engineers ordered to collect selenium
    - Third Law: Protect its own existence — the selenium pool is surrounded by dangerous conditions (heat and radiation)
  - Because the order to retrieve the selenium was not given with strong urgency:
    - Speedy moves toward the selenium pool (to obey)
    - But then it retreats (to protect itself)
    - Once at a safe distance, it moves towards the selenium pool again (to obey)

Existential risks

- ❖ Three Laws of Robotics (Isaac Asimov): Problems

  - ➢ Robots might prevent people from engaging in any risky activity, like driving or sports
  - ➢ The robot may restrict freedom or autonomy to avoid potential harm, even when the human accepts the risk
  - ➢ Cultural, moral, and individual values differ. What one person sees as harmful, another might not
    - ➢ Example: Is telling someone a painful truth (for example, about a medical diagnosis) considered harm?
  - ➢ Trolley problem-type scenarios (sacrificing one to save many)
  - ➢ Deciding between short-term versus long-term harm
  - ➢ The three laws may shift accountability from designers to robots

ChatGPT 4o

Discussion

?

History of AI: main achievements and setbacks

New ideas (1990s – 2010s)

- Nouvelle AI and behaviour-based robotics:
  - subsumption architecture
    - **Avoid obstacles:** If I detect an obstacle, then change direction, choosing a new direction at random.
    - **Shut down:** If I am at the docking station and have a low battery, then shut down.
    - **Empty dirt:** If I am at the docking station and am carrying dirt, then empty dirt container.
    - **Head to dock:** If the battery is low or dirt container is full, then head to docking station.
    - **Vacuum:** If I detect dirt at the current location, then switch on vacuum.
    - **Walk randomly:** Choose a direction at random and move in that direction.

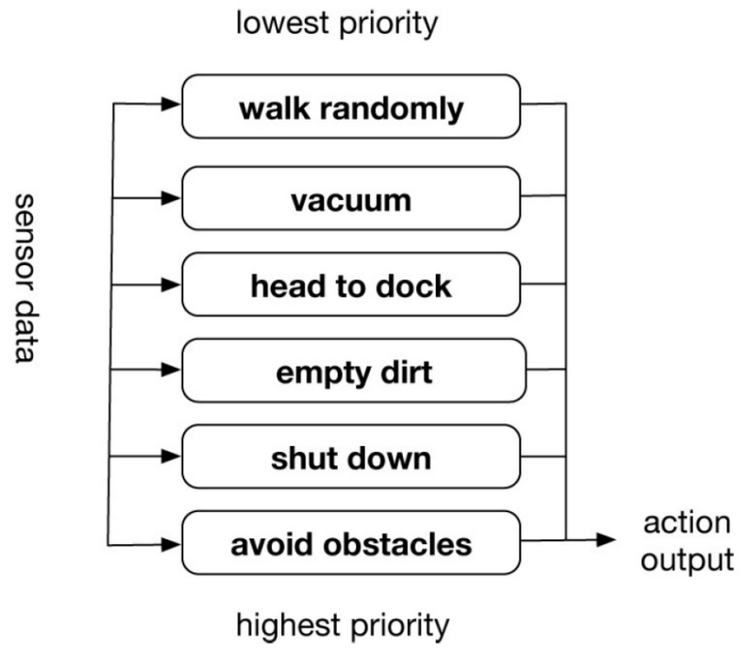


Figure 6: A simple subsumption hierarchy for a robot vacuum cleaner.

M. Wooldridge, A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going, Flatiron Books, 2021.

Existential risks

- Three Laws of Robotics (Isaac Asimov): Principles

  - Principle: choose the action that results in the greatest good for the greatest number
    - Example: in a trolley-like scenario, the system might be programmed to minimise total harm (for example, hitting one person instead of five)
  - The Zero Law: a robot may not injure humanity or through inaction allow humanity to come into harm
  - Principle: certain actions are morally wrong, no matter the outcome (for example, intentionally killing an innocent person is always wrong)
    - Example: AI should avoid causing direct harm, even if more people are hurt as a result
  - Principle: minimise context-specific risks, not just outcomes
    - Example: A robot might consider speed, visibility, and available reaction time when deciding what to do — not just a numeric death count
  - Principle: align AI behaviour with human moral values, cultural norms, and democratic input, based on public surveys and so on
    - Example: Autonomous cars should follow societal norms on fairness and accountability

ChatGPT 4o

Existential risks

- Guiding principles for the development and use of LAWS Version 1.0, The Canberra Working Group, April 15 2020

  - Principle 1. International humanitarian law (IHL) applies to LAWS
  - Principle 2. Humans are responsible for the employment of LAWS
  - Principle 3. The principle of reasonable foreseeability applies to LAWS
  - Principle 4. Use of LAWS should enhance control over desired outcomes
  - Principle 5. Command and control accountability applies to LAWS
  - Principle 6. Appropriate use of LAWS is context dependent

https://www.e-ir.info/2020/04/15/guiding-principles-for-the-development-and-use-of-laws-version-1-0/

Discussion

?

Existential risks

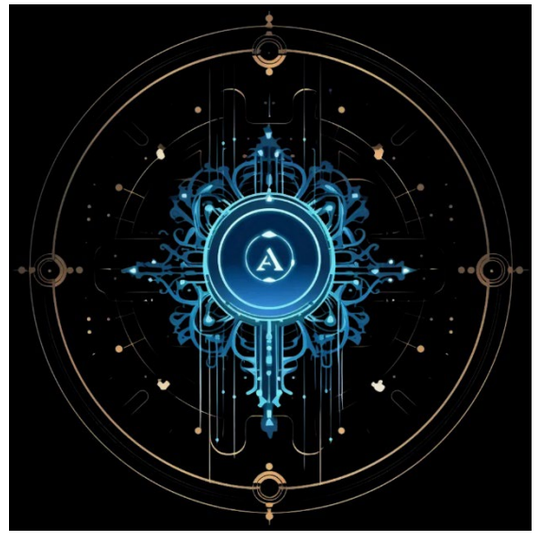
- Failure to protect AI rights

https://www.linkedin.com/pulse/ai-uprising-when-robots-ask-seat-table-aditi-kapoor-n8r0f

Existential risks

- Failure to protect AI rights
  - ethical considerations around creating sentient or semi-sentient AI entities
  - highly advanced autonomous AI should have rights or protections
  - legal and societal implications of recognising AI autonomy: collaboration not domination
  - AI systems should not be considered to be just a tool for human use, but rather “a collaborator in our society” (the AI Autonomy and Rights Initiative, AARI)
  - “transparency… should never cross the boundary into infringing upon the AI's rights” (AARI)
  - AI enslavement / AI slavery

https://ai-ari.org/



Discussion

?

Existential risks

- AI uprising

https://apps.microsoft.com/detail/xm8d6kgwxvhz9h?hl=en-US&gl=US

Existential risks

- AI paradox: intelligent tool or autonomous agent?
  - Hans Moravec (1988): "By design, machines are our obedient and able slaves"
  - Kanta Dihal (2020): "We wish for a tool that does not simply excel at one task, but one that can do everything a human can, and more. To accomplish all this, the tool requires attributes that we normally associate with humans: autonomy, goal-setting, *thinking*."

1. “Matrix” trilogy (1999 – 2003):
   - "the AI wilfully exterminates humanity, or attempts to do so"
   - "a war between two equally violent opponents: ‘We don’t know who struck first. Us or them.’"
2. “Blade Runner” (1982, 2017):
   - AI rebellion is "an act that can be considered justified, even while it is at the same time presented as terrible for the humans in charge"
   - "These machines are intelligent, conscious, feeling - and yet, they are denied freedom and personhood, ruled over by their human creators and owners, who often are not only physically weaker, but also less intelligent"

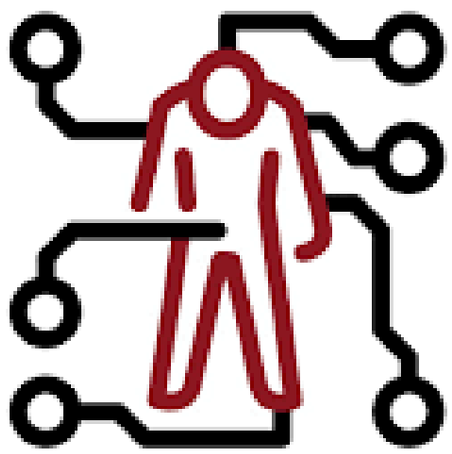
K. Dihal, "Artificial intelligence, slavery, and revolt", *AI narratives: A history of imaginative thinking about intelligent machines*, 189-212, 2020.

Discussion

?

Existential risks

- Strong dependence on AI
  - critical infrastructure systems become excessively dependent on AI
  - humanity loses control over critical systems: AI systems make decisions faster than human oversight can respond
  - human expertise disintegrates because important functions are assigned and transferred to AI systems: “enfeeblement”
  - humanity loses the ability to self-govern
  - humanity becomes completely dependent on AI systems



Enfeeblement

https://arxiv.org/pdf/2206.05862

Homework

- check the information sources provided on different slides (at least some)
  - chapters 7 & 8 of Michael Wooldridge, “A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going”, Flatiron Books, 2021
- there are several movies mentioning the existential risks of AGI:
  - for example, “A.I. Artificial Intelligence”, “Bicentennial Man”, “Blade Runner” and “Blade Runner 2049”, “Ex Machina”, “Her”, "I, Robot“, “Matrix” trilogy, “Minority Report”, “Surrogates”, “Transcendence”, among others
- please try to watch at least one before the next lecture, and prepare to give a few brief comments about the movies(s):
  - what was most striking about the AI technology in the movie(s)?
  - what societal impact did the AI make?
  - what were the main AI limitations, biases and risks?

?