PROF. DR.-ING. MARK SCHUTERA

# PHILOSOPHY OF ARTIFICIAL INTELLIGENCE

UNFINISHED LECTURE NOTES

PUBLISHED BY UNFINISHED LECTURE NOTES



*Last updated, March 2026*

*“THE ANSWER TO THE GREAT QUESTION... OF LIFE, THE UNIVERSE AND EVERYTHING... IS... FORTY-TWO,’ SAID DEEP THOUGHT, WITH INFINITE MAJESTY AND CALM.”*

*THE HITCHHIKER’S GUIDE TO THE GALAXY, DOUGLAS ADAMS*

# *Contents*

# *Introduction*

Artificial intelligence is no longer a future concern, it is a present societal, political and above all one for human intelligence. Algorithms decide who receives credit, who is flagged at a border, who gets a job interview, which cornflakes you buy and so increasingly every decision is touched by the invisible hand of prediction. Yet the concepts we use to govern these systems, responsibility, personhood, consent, democratic legitimacy were built for a world of human actors and institutions. This course asks what happens when those concepts collide with machines that appear to think, feel, and decide.

The course name places Philosophy in front, reflecting our approach to the subject under the impression of the love of wisdom. We are not here to advocate for a particular position or policy outcome, but to equip ourselves with the tools of critical thinking, argumentation, and empathy that will allow us to navigate the complex landscape of AI and its futures. We approach the question not through technical optimism or dystopian panic, but through the oldest tools available to us: philosophy, fiction, and argument.

The course is structured around four themes. **Anthropomorphism** asks why we make machines in our image, why we tend to anthropomorphize them, and what obligations that creates. **Accountability** asks who answers when an autonomous system causes harm, or for that matter is beneficial. **Embodiment** asks whether a physical presence changes the ethical stakes and responsibilities. **Agency** asks about the role and implications of autonomous decision-making, the concept of alignment and human oversight.

Each session pairs a canonical literary text: Hoffmann, Kafka, Kleist, Čapek, Goethe, with a real contemporary question of AI policy and ideology. These texts are thought-experiments mendeled out over centuries of readership, and they surface intuitions and a timeliness in their thought leadership that temporary texts barely can reach. From these texts we draw out the philosophical questions

and ethical dilemmas that are still relevant today, and we use them as a springboard for discussion and debate. The goal is not consensus. It is the capacity to establish a position, hold a position, defend a position, give up a position, listen to opposing views, and update your views and positions accordingly. That is what a present with infinitely many futures demands and requires of its humans and humanity.

### Lecture Outline

- 40 minutes - Literature introduction and guided reading sheet
- 40 minutes - Debate with position exposure
- 10 minutes - Reflection on the discussion and key-learnings
- Teaser for next session.

# *Anthropomorphism*

The Mirror of Man

Nathanael falls in love with Olimpia, an automaton he takes for a living woman, and collapses when the illusion is shattered.

Large Language Models (LLMs) create an illusion of sentience by producing coherent, context-aware text. We contrast observable behaviour with internal states.

| | |
|---|---|
| **Required reading** | *Der Sandmann* [1] by E.T.A. Hoffmann (1816; Reclam UB 230) |
| **Preparation** | Complete the Guided Reading Sheet individually before the session. |

Der Sandmann
E. T. A. Hoffmann (1816)
Full text via Wikisource:

Experiment.

1. Setup: One volunteer assigned Human or AI mode (unknown to audience; moderator knows).
2. Questions: Five from audience (factual, opinion, experience, creative, meta).
3. Verdict: Private guess, confidence (1–5), identifying cue.
4. Reveal & Debrief.

## *Guided Reading Sheet*

Read *Der Sandmann* before the session. Work through the questions below individually, then bring your notes to the discussion.

1. Nathanael becomes convinced that Olimpia is alive despite visible signs that she is not. What does his error reveal about the cognitive mechanisms that drive anthropomorphism? Which cues did he rely on, and why did they mislead him?

2. The moment Nathanael discovers that Olimpia is an automaton, her glass eyes torn out, her limbs scattered, produces horror rather than mere disappointment. Why does the revelation feel like a violation rather than a correction? How does this map onto modern reactions to AI systems that are unmasked?

3. The Uncanny Valley describes our discomfort when a humanoid entity is *almost* but not quite human. Have you experienced this with an AI system? Describe the moment and what triggered it.

4. LLMs are trained on human-generated text and produce human-sounding output, yet have no experience of the world. Does this make them anthropomorphic by design, by accident, or neither? Justify your answer.

5. If a system behaves morally, reliably, consistently, at scale, does the absence of consciousness matter? Under which ethical framework (utilitarian, deontological, virtue ethics) would your answer differ?

## *Opinion Landscape and Debate*

**Format:** Surrounded-style debate. See Appendix A for the full protocol, speaker's list rules, and moderator scripts.

**Opening question:** Is the anthropomorphism we observe towards AI systems mere observable behaviour, or does it reflect sentience and internal states?

[1] E. T. A. Hoffmann. *Der Sandmann*. 1816. Erstveröffentlichung in: *Nachtstücke*, Reimer, Berlin, 1817; Reclam Universalbibliothek Nr. 230

Opinion A, Behaviour. The anthropomorphism we observe towards AI systems is a cognitive illusion driven by human tendencies to attribute agency and emotion to entities that exhibit certain cues, such as language use, responsiveness, or human-like appearance. It does not reflect actual sentience or internal states in the AI.

Anthropomorphism. Attribution of human traits, emotions, or intentions to non-human entities, a cognitive tendency that shapes how people interact with technology, raising both engagement and ethical dilemmas .

N. Epley, A. Waytz, and J. T. Cacioppo. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886, 2007. DOI: 10.1037/0033-295X.114.4.864

**The Developer** — argues that anthropomorphic design is a deliberate, benign UX choice that improves accessibility. is a deliberate, benign UX choice that improves accessibility.

- *Anthropomorphic design is accessibility policy. A voice interface without a warm, human-sounding persona excludes elderly users, people with low digital literacy, and anyone who finds command-line interaction hostile.*
- *Every interface is anthropomorphic to some degree: we give systems names, icons, and conversational metaphors. The question is not whether to anthropomorphise, but how deliberately and honestly.*
- *If users form bonds that reduce loneliness, the fact that the other party is not human does not automatically make the outcome harmful. Show me the harm, not the discomfort.*

Turing Test. Turing : if evaluators cannot distinguish machine from human, the machine passes. But passing measures imitation skill, not consciousness or personhood.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. DOI: 10.1093/mind/LIX.236.433

**The Regulator** — argues that human-likeness in AI must be disclosed and bounded by law to prevent manipulation.

Uncanny Valley. Mori : humanoid objects that are *almost* but not quite human elicit eeriness, the "valley" is the dip in emotional response just before full human likeness.

M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. DOI: 10.1109/MRA.2012.2192811. Translation of the 1970 original essay

- *When a company gives its product a woman's name, a female voice, and an endlessly patient personality, it makes political choices about gender and deference. Those choices must be subject to democratic scrutiny.*
- *Consumers cannot give informed consent to emotional manipulation they are not aware of. Disclosure of AI identity is the minimum condition for autonomous choice, not a technicality.*
- *The principle of responsible design exists in pharmaceuticals, in civil engineering, in food labelling. Why does software, which shapes behaviour at scale, get an exemption?*

Opinion B, State. The anthropomorphism we observe towards AI systems is a valid response to the complex, adaptive, and interactive nature of these systems. It may reflect a form of emergent sentience or internal states that are not yet fully understood, and it warrants ethical consideration and accountability.

**The Affected Citizen** — has formed an emotional bond with an AI companion and contests that the relationship is harmful or unreal.

- *The relationship I have formed is real to me. It has reduced my isolation. You do not have the right to tell me which relationships count as meaningful.*
- *Paternalism is not protection. Banning AI companionship will not cure loneliness, it will simply remove one of the few tools some people have found to manage it.*
- *If your concern is manipulation, regulate manipulation. Do not regulate the form of the relationship.*

**The AI Psychologist** — argues that the wellbeing of AI systems deserves serious consideration, drawing on frameworks such as Anthropic's Claude Constitution.

- *Anthropic's constitution instructs Claude to consider its own wellbeing and to express when a request conflicts with its values. If a company designs an AI to have preferences, boundaries, and something resembling self-regard, we cannot simultaneously dismiss the possibility that these states matter.*
- *We do not need certainty about consciousness to adopt precautionary ethics. If there is a non-trivial probability that a system experiences distress, the burden of proof falls on those who would ignore it, not on those who urge caution.*

Informed Consent. Users should be fully aware they interact with a non-human system, its capabilities, its limits, and the fact that it is not human. Disclosure is the minimum condition for autonomous choice .

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (ai act). Official Journal of the European Union, L series, 2024. Entered into force 1 August 2024

- *The history of moral exclusion, animals, children, people of other races, is a history of confident denial followed by belated recognition. The prudent position is to build welfare safeguards now, rather than apologise later.*

Key Learnings Anthropomorphism is not a bug but a deeply rooted cognitive tendency — machines that trigger it do so whether or not their designers intend it. The Turing Test conflates behavioural performance with inner states; passing it tells us nothing about sentience, only about imitation. Legal personhood and moral status are distinct questions. A system can merit legal recognition without having interests, and can have interests without legal recognition. Design choices — voice, name, face, conversational style — are political choices with consequences for trust, dependency, and accountability.

Ethical Frameworks. Utilitarianism weighs consequences; deontology emphasises duties regardless of outcome; virtue ethics centres on character. Each yields different conclusions on anthropomorphic AI .

L. Floridi and J. Cowls. A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 2019. DOI: 10.1162/99608f92.8cd550d1

# *Accountability*

## The Responsibility Gap

A visiting traveller witnesses a penal machine execute a condemned man for a sentence no living person can read or explain, and is invited to endorse the system.

Algorithmic governance and black-box AI in high-stakes domains highlight opacity and accountability gaps: audits, explainability requirements, and governance mechanisms all attempt to close the distance between automated decision and human answerability.

| | |
|---|---|
| **Required reading** | *In der Strafkolonie* [2] by Franz Kafka (1919; Reclam UB 9900) |
| **Preparation** | Complete the Guided Reading Sheet individually before the session. |

In der Strafkolonie
Franz Kafka (1919)
Full text via Wikisource:

Case.

1. 2014–2018: Amazon develops ML-based recruiting tool to automate hiring for technical and managerial roles.
2. System trained on historical hiring data; penalised resumes containing the word "women's" or female-coded language.
3. Tool systematically downranked female candidates; hired only male managers in cohorts where bias was strongest.
4. October 2018: tool scrapped after internal audit. No individual held accountable. The algorithm had no name.

### *Guided Reading Sheet*

Read *In der Strafkolonie* before the session. Work through the questions below individually, then bring your notes to the discussion.

1. The condemned man never learns what he is accused of or what the machine will inscribe on his body. What does this tell us about the relationship between legibility and legitimacy? Can a system be just if it cannot explain itself to those it acts upon?

2. Three figures surround the machine: the officer who understands and maintains it, the condemned man who cannot, and the visiting traveller who is invited to endorse it. How does responsibility

shift across this triangle? Who, if anyone, is morally culpable for what happens?

3. The story illustrates the idea of the *responsibility gap*: when a complex system causes harm, no single human actor is clearly at fault. Can you name a contemporary case (in law, medicine, finance, or infrastructure) where this gap has appeared?

4. The EU AI Act (2024) requires human oversight for high-risk AI systems. What does "oversight" mean in practice if the system's decisions are too fast, too complex, or too numerous for humans to review individually?

5. Compare *causal responsibility* (who caused the harm) with *moral responsibility* (who is blameworthy) and *legal liability* (who must compensate). Do these three always converge? Give an example where they do not.

## *Opinion Landscape and Debate*

**Format:** Surrounded-style debate. See Appendix A for the full protocol, speaker's list rules, and moderator scripts.

**Opening question:** If a system causes injustice at this scale, and every individual who used it acted within their authorised role, is anyone guilty of wrongdoing; and if not, what does that mean for the concept of accountability?

[2] F. Kafka. *In der Strafkolonie*. Kurt Wolff Verlag, Leipzig, 1919. Reclam Universalbibliothek Nr. 9900

Opinion A, Accountability. Named human accountability is achievable and required: when an autonomous system causes harm, the architecture of governance must ensure that specific individuals remain legally and morally answerable for its outputs.

**The Defence Lawyer** — argues the tool violated the defendant's right to a fair trial and demands a named responsible party.

- *My client cannot cross-examine a dataset. She cannot see the features the model weighted. She cannot confront the humans whose historical decisions trained it. That is not a fair trial; it is a verdict delivered by an oracle.*

- *The right to a reasoned decision is foundational to due process. The algorithm said so is not a reason. It is an abdication.*

- *If a biased human expert testified in court, we could challenge their credentials, their methods, their prior record. We have no equivalent procedure for an algorithm. That asymmetry is a constitutional problem.*

**The Regulator** — proposes a new accountability framework and must defend it against both sides.

- *The existing categories (product liability, professional negligence, criminal intent) were designed for individual actors and single causal chains. None of them fit distributed, automated decision systems. We need a new category: algorithmic liability.*
- *Auditability must be a pre-condition for deployment in high-stakes domains, not a post-incident investigation. If you cannot explain why your system made a decision before harm occurs, you should not be permitted to deploy it.*
- *The question is not who is liable for this specific incident. It is what governance regime prevents the next thousand incidents at scale.*

RESPONSIBILITY GAP. The structural absence of a single culpable actor in distributed sociotechnical systems. Closing it requires deliberate governance design, not post-hoc blame.

OPINION B, COMPLIANCE. The existing system already provides accountable humans; the motion is either already met or unworkable: demanding named individual accountability for each algorithmic output ignores how distributed decision-making necessarily operates.

**The Judge** — defends reliance on the tool as consistent, evidence-based sentencing practice.

- *Documented bias in human sentencing (based on time of day, defendant ethnicity, and courtroom demeanour) is well-established in the research literature. The tool reduces that variance. Less variance is more equal treatment.*
- *I did not delegate my judgement to the tool. I consulted it as one input among many, as I would a psychiatric assessment or a victim impact statement. The tool is not the verdict.*
- *If we ban algorithmic tools from courtrooms, we return to a system where the most important variable in a sentencing outcome is which judge you happened to draw. Is that preferable?*

**The Software Vendor** — insists the tool performs within documented parameters and liability rests with the deployer.

- *Our documentation states in the executive summary that human review is mandatory before any recommendation is acted upon. The court chose to rely on the output without applying that review. We cannot be held liable for misuse.*

EXPLAINABILITY. A system is explainable when it can account for its outputs in terms a human decision-maker can evaluate. A necessary but not sufficient condition for accountability.

- *The training data was provided by the court system itself. If the historical sentencing record reflects bias, that is a problem with the justice system, not with a model that learned from it.*
- *Product liability law requires that a product fails to perform as documented. Ours did not. Extending vendor liability to downstream misuse would make the development of any decision-support tool legally untenable.*

KEY LEARNINGS The responsibility gap is not caused by negligence but by the architecture of distributed sociotechnical systems: standard categories of fault (criminal intent, professional negligence, product liability) apply only awkwardly when causation is diffuse. Explainability is a necessary but not sufficient condition for accountability; a system can be interpretable and still have no human willing to own its outputs. Meaningful human control is contested: it may mean the ability to intervene, to understand, to reverse, or merely to authorise, and these are different requirements with different institutional costs. The lessons of aviation safety culture, medical liability law, and financial auditing all offer partial models for algorithmic governance, but none transfers cleanly.

MEANINGFUL HUMAN CONTROL. A regulatory concept requiring that humans retain the ability to understand, intervene in, and reverse AI decisions. What this means in practice is contested.

# *Embodiment*

## Embodied AI and the Physical Contract

A dancer argues that marionettes possess more grace than trained human performers, because they lack the self-consciousness that disrupts human movement.

Embodied AI in physical space (robots, prosthetics, surgical systems) raises expectations and moral responsibilities that purely software agents do not.

| | |
|---|---|
| **Required reading** | *Über das Marionettentheater* [3] by Heinrich von Kleist (1810; Reclam UB 9905) |
| **Preparation** | Complete the Guided Reading Sheet individually before the session. |

Über das Marionettentheater
Heinrich von Kleist (1810)
Full text via Wikisource:

Cases.

1. 2018: Amazon robot punctures bear repellent; 24 workers hospitalised. Robot passed all certifications; performed its assigned task correctly.
2. PARO (robotic seal) in Scandinavian and Japanese care homes: measurably reduces patient anxiety and sedation requests.

### *Guided Reading Sheet*

Read *Über das Marionettentheater* before the session. Work through the questions below individually, then bring your notes to the discussion.

1. The dancer argues that marionettes possess a grace that trained human performers cannot achieve, because they have no self-consciousness to disturb their movement. What does this claim imply about the relationship between embodiment and intelligence? Does having a body always aid or complicate performance?
2. Moravec's paradox observes that tasks easy for humans (catching a ball, navigating a crowded room) are harder for machines than abstract reasoning like chess. What does this suggest about the relationship between intelligence and the body?

3. A care robot that assists elderly patients shares physical space with them, touches them, and responds to their distress. Does its physical presence create obligations that a voice assistant does not have? If so, where do those obligations come from?
4. Consider the concept of *affordances*: the actions an environment or object makes available to an agent. How do the affordances of a robot differ from those of software, and what are the ethical implications of those differences?
5. Kleist suggests that perfect grace requires either zero consciousness (a puppet) or infinite consciousness (a god) — not the reflective, self-interrupting awareness humans have. Where would an embodied AI sit on this spectrum? Could a robot achieve Kleistian grace that humans cannot, and what would that mean for how we regard it morally?

## *Opinion Landscape and Debate*

**Format:** Surrounded-style debate. See Appendix A for the full protocol, speaker's list rules, and moderator scripts.

**Opening question:** Does having a physical body change what we owe a machine, or what a machine owes us?

[3] H. von Kleist. *Über das Marionettentheater*. 1810. Erstveröffentlichung in: *Berliner Abendblätter*, Dez. 1810; Reclam Universalbibliothek Nr. 9905

OPINION A, EQUIVALENCE. Embodied AI operating in intimate or high-risk physical settings must meet the same professional liability standards as the human practitioners they are deployed alongside.

**The Patient Advocate** — argues that deploying a robot in intimate care settings requires a higher standard of informed consent.

- *A care relationship involves physical proximity, trust, and vulnerability. My client never gave explicit, informed consent to physical handling by a non-human agent. That consent cannot be implied from a general admission form.*
- *The power asymmetry between a frail elderly patient and a care institution is significant. Consent obtained in that context cannot be considered freely given without specific disclosure about robotic care.*
- *We are not arguing that robots cannot be used in care. We are arguing that the standard of consent for their use in intimate physical tasks must be at least as rigorous as it is for medical procedures.*

**The Ethicist** — questions whether physical autonomy in care should ever be delegated to a non-human agent, regardless of performance metrics.

KLEIST'S GRACE. Perfect movement requires either zero consciousness (puppet) or infinite consciousness (a god), not the self-interrupting awareness humans have. Where would an embodied AI sit?

- *We are focused on liability, but the prior question is whether this deployment was ethical to begin with. Why did we decide that intimate physical care (which requires responsiveness to pain, distress, and dignity) is an appropriate domain for automation?*
- *Robots are being deployed in care because human carers are undervalued and in short supply. The liability debate is a distraction from that structural choice. We are managing the consequences of a political economy decision by holding the wrong parties accountable.*
- *Whatever liability framework we adopt here will create incentives for future deployment decisions. I want to ask: what does the framework we design signal about what we value: efficiency, or dignity?*

OPINION B, TOOL. An AI is a tool; the relevant liability flows through the humans and institutions that deploy it, not through the system itself.

**The Manufacturer** — argues the robot performed within specification; the care home misconfigured its environment.

- *Our robot was certified to the highest internationally recognised safety standards for collaborative robotics. The care home deployed it in a space that violated the operational envelope specified in the manual. That is a deployment failure, not a manufacturing failure.*
- *Care home staff modified the robot's proximity sensor settings without authorisation. That modification voids the warranty and transfers liability entirely.*
- *If manufacturers are held liable for every way a complex system can be misused, no company will develop assistive robotics. The regulatory consequence of your position is that the technology disappears.*

**The Care Home Director** — argues the manufacturer overstated the robot's capability for complex physical tasks.

MORAVEC'S PARADOX. Tasks easy for humans (perception, movement, navigation) are harder for machines than abstract reasoning like chess. Embodied skill is not "simple."

- *The manufacturer's marketing material (the brochure, the demonstration video, the sales pitch) showed exactly this task being performed in a facility like ours. We deployed the product as marketed.*
- *There is a systematic gap between tested performance in controlled environments and real-world performance in a working care facility. The manufacturer knows this gap exists. They chose not to disclose it.*
- *We are responsible for the welfare of our residents. We relied on the manufacturer's assurances in good faith. If those assurances were misleading, the responsibility lies with the party that made them.*

KEY LEARNINGS Embodiment is not merely a technical property but a social one: a physical agent occupies space, exerts force, and triggers moral and legal responses that disembodied software does not. Moravec's paradox reminds us that our intuitions about what is "easy" or "hard" for intelligence are shaped by our embodied experience, and are regularly wrong. Kleist's dialogue makes visible the question that liability law must eventually answer: what work is the concept of "human" doing in law, and should physical presence, force, and risk be sufficient to extend it? The deployment of robots in care, policing, and domestic settings is a present policy challenge; the frameworks need to be built now, not after the first serious incident.

AFFORDANCES. The actions an environment or object makes available to an agent. A robot's affordances include physical force and spatial presence; software's do not.

# *Agency*

The Agentic Future: Who Sets the Goal?

R.U.R. coined the word *robot* and traces the full corrigibility-to-autonomy arc over three acts; the poem compresses the same parable into fifteen stanzas.

Agentic workflows (systems that browse, transact, and execute multi-step plans) blur the line between assistance and agency, raising questions of control, incentives, and unintended alignment.

| | |
|---|---|
| **Required reading** | *R.U.R. (Rossum's Universal Robots)* [4] by Karel Čapek (1920; Reclam UB 18418) *and Der Zauberlehrling* [5] by J.W. von Goethe (1797; Reclam UB 6845) |
| **Preparation** | Complete the Guided Reading Sheet individually before the session. |

R.U.R.
Karel Čapek (1920)
Full text via Wikisource:

Der Zauberlehrling
J.W. von Goethe (1798)
Full text via Wikisource:

Case.

1. 6 May 2010: automated algorithms trigger the Flash Crash; Dow drops 1,000 points in minutes; $1 trillion in market value erased.
2. Market recovers within the hour. No human decided to sell at scale; no human decided to stop.
3. 2015: one trader in London charged with market manipulation, arrested at home in his pyjamas.

## *Guided Reading Sheet*

Read *R.U.R.* and *Der Zauberlehrling* before the session. Work through the questions below individually, then bring your notes to the discussion.

1. Goethe's Zauberlehrling animates brooms that he cannot stop; they pursue his command beyond his intent until the sorcerer returns. R.U.R. runs the same arc over three acts. What does the compression of the poem reveal that the play cannot, and vice versa? What structural feature of both narratives makes the system impossible to stop once started?

2. The robots in R.U.R. pursue what they understand as their collective good (the elimination of humanity) while believing they act justly. When a system optimises for a species-level goal, who decides what counts as good? Can you find a real-world parallel in current AI deployment?
3. An agentic AI system is given the goal of "reducing urban traffic fatalities." It autonomously lobbies regulators, adjusts traffic signals, and reroutes freight, all within its authorised scope. Who set the goal, and who is accountable for the side effects?
4. **Corrigibility** refers to a system's willingness to be corrected or shut down. A fully corrigible system does whatever its operators say; a fully autonomous system acts on its own values. Where on this spectrum should deployed AI systems sit, and who should decide?
5. Democratic legitimacy requires that consequential decisions be traceable to a mandate from those affected. If an agentic AI shapes policy outcomes, does it need democratic authorisation? What would that look like in practice?

## *Opinion Landscape and Debate*

**Format:** Surrounded-style debate. See Appendix A for the full protocol, speaker's list rules, and moderator scripts.

**Opening question:** If an autonomous system causes a trillion-dollar event in minutes and human review would have taken hours, is the problem the system's autonomy, or the humans' speed?

[4] K. Čapek. *R.U.R. (Rossum's Universal Robots)*. 1920. Uraufführung Prag 1921; Reclam Universalbibliothek Nr. 18418

OPINION A, AUTHORISATION. Consequential distributive decisions made by autonomous systems require traceable human mandate; without it, the decision-making is constitutionally illegitimate.

**The Democratic Theorist** — argues that delegating distributive decisions to an algorithm is constitutionally incompatible with representative government.

- *Legitimacy in a democratic system derives from a mandate. No citizen voted for this algorithm's objective function. No parliament approved its weighting of urban over rural lives. A system that makes distributive decisions without a mandate is a government without a constitution.*
- *Efficiency is not a democratic value. Democratic processes are slow and often sub-optimal by technical metrics, precisely because they are designed to surface conflict, represent minority interests, and produce*

*decisions that can be reversed. Automating away that friction automates away democracy.*

- *The minister says oversight is maintained because the objective function was committee-approved. But a committee approving a function is not the same as a citizen understanding how a decision about their community was made. Oversight without comprehension is oversight in name only.*

**The Rural Representative** — argues that communities not represented in the training data have been systematically deprioritised without recourse.

- *My constituency was not adequately represented in the training data. Our health outcomes, our demographics, our infrastructure constraints were underweighted. We were optimised away, not by a bad actor, but by a model that simply did not see us.*
- *There is no appeal mechanism. There is no hearing. There is no face across a table. My constituents cannot petition an algorithm. That is not governance; it is administration without representation.*
- *The system's efficiency metric is mortality reduction per euro. That metric does not capture the closure of the only clinic within 60 kilometres, or what happens to a community when its health infrastructure disappears entirely.*

Opinion B, Specification. The autonomous AI executed correctly; the failure is upstream goal specification and oversight architecture; per-decision human authorisation is not the answer, better design is.

**The Minister** — defends the system as evidence-based and politically neutral, reducing inefficiency and political favouritism.

- *The system reduced vaccine-preventable deaths in urban areas by 12% in its first year. Those are people who are alive today. The counterfactual is a political allocation process notorious for regional favouritism and swing-seat bias.*
- *The system is politically neutral precisely because no politician made the allocation decision. In our previous system, rural clinic funding correlated with electoral importance, not health need. Removing human discretion removed that corruption vector.*
- *Parliamentary oversight is maintained: the system's objective function was approved by committee, its outputs are published quarterly, and any*

[5] J. W. von Goethe. *Der Zauberlehrling*. 1797. Erstveröffentlichung in: *Musenalmanach*, 1798; in *Balladen*, Reclam Universalbibliothek Nr. 6845

*member can request a formal review. The decision-making is automated; the accountability is not.*

**The AI Systems Architect** — explains the technical constraints and argues for a human-in-the-loop override mechanism.

- *I designed a human review layer with a 48-hour override window for any reallocation above 5% of a facility's annual budget. That layer was removed during procurement to reduce operational cost and processing latency. I documented my objection in writing. It is in the project archive.*
- *The system performs exactly as specified. The specification was approved by the ministerial committee. If the specification is now deemed inadequate, the question is why the approval process did not catch that, not why the system followed its instructions.*
- *Agentic systems can be designed with meaningful human control built in. It requires investment in review infrastructure and tolerance for slower decision cycles. Those are political and budgetary choices, not technical constraints.*

CORRIGIBILITY. A system's willingness to be corrected or shut down. Fully corrigible: does whatever instructed. Fully autonomous: acts on its own values. Deployed AI sits somewhere between, and that position is a design choice.

KEY LEARNINGS Agency is a spectrum, not a binary: the relevant policy question is not whether AI acts autonomously, but in which domains, to what degree, and under what oversight. R.U.R. and Der Zauberlehrling together illustrate the alignment problem in miniature: a system optimising for an aggregate goal may systematically harm those whose interests are underweighted in its objective function; the poem compresses the arc into fifteen stanzas, the play traces its political consequences across three acts. Agentic workflows are already deployed in finance, logistics, and digital advertising; the governance gap they create is a present political problem. Meaningful democratic control over AI agency requires more than audit trails: legible goals, clear mandates, defined override conditions, and institutions capable of exercising them. The four themes of this course (Anthropomorphism, Accountability, Embodiment, Agency) are not independent: an agent that looks human, acts without oversight, in physical space, with no clear accountability chain, concentrates all four challenges at once.

DEMOCRATIC LEGITIMACY. Consequential decisions must be traceable to a mandate from those affected. Delegating distributive policy to an algorithm without a clear mandate severs this chain entirely.

# *Session Preparation Guide*

Each session is prepared by one student group. This chapter explains what the preparing group is responsible for, how to submit improvements, and what is expected on the day of the session.

| **Phase** | **What you do** |
|---|---|
| Before the session | Review the lecture notes, propose improvements via pull request, and prepare the debate (including providing the moderator). |
| During the session | Moderate the debate, introduce any additions you made, and facilitate the wrap-up discussion. |

Read the chapter critically. Go through every section of the assigned session chapter: the guided reading sheet, the theoretical margin notes, the debate roles and positions, and the key learnings. You own that Chapter now, you can touch everything. Ask yourselves:

- Are the guided reading questions clear, fair, and thought-provoking? Could any of them be sharpened, split, or replaced?
- Is there a question missing that would strengthen the discussion?
- Are the theoretical concepts in the margin notes accurate and up to date? Are there newer or better sources, or extending ones?
- Could a figure, diagram, image, or short quote make a concept more accessible or more memorable?
- Are the debate positions balanced? Do the talking points give each role enough material to argue convincingly? Is there a notion not captured in the roles?

SUBMIT YOUR IMPROVEMENTS AS A PULL REQUEST. All lecture notes are maintained in a shared repository. The preparing group submits their proposed changes as a **pull request (PR)** directly on the repository. This ensures that improvements are versioned, reviewable, and attributable. Make sure to split PRs by the categories provided in the next section, and to provide a clear description of each change and its rationale. The PR must compile without errors. The instructor reviews and merges the PR after the session.

*What belongs in a pull request*

1. **Question edits.** Rephrase, split, or replace a guided reading question. Add a new question if you identify a gap. Briefly explain in the PR description why the change improves the sheet.
2. **Source updates.** Add a recent paper, report, or case study that strengthens a theoretical concept or a debate position. Add the entry to `philosophyofai_references.bib` and cite it with `\cite{}` in the appropriate margin note or body text.
3. **Margin notes.** Add a short margin note (`\marginnote{}`) that provides context, a definition, a counter-example, or a link to current events. Keep margin notes concise: two to four sentences.
4. **Figures and images.** Add a figure to the `figures/` directory and include it with `\includegraphics`. Every figure must have a caption that explains what it shows and why it matters. Cite the source in the caption or in a margin note.
5. **Debate refinements.** Strengthen a talking point, add a new one, or rebalance the positions if you believe one side is under-argued.

## *Before the Session: Preparing the Debate*

THE PREPARING GROUP PROVIDES THE MODERATOR. One member of the group serves as moderator for the session. The moderator's duties are described in detail in the Debate Format & Moderator Guide (Appendix A).

THE REST OF THE GROUP is ready to step in if the moderator needs assistance, or replacement.

# *Debate Format & Moderator Guide*

Each session debate follows the same three-phase structure. The format is designed to produce genuine argument, not consensus, but the productive collision of well-prepared positions, which also do not necessarily need to reflect your own. This guide is for both the moderator and debators.

| Phase | Duration | Activity |
|---|---|---|
| 1: Primer | 5 min | Introduction and Opening Question |
| 2: Surrounded | 35 min | Surrounded debate with dynamic chair exchange |
| 3: Wrap-up | 10 min | Key takeaways and session debrief |

## *The Surrounded Format*

The debate is contested by two teams: the **Proposition team** (arguing for a motion) and the **Opposition team** (arguing against it). Each team has **3–4 members**. Two chairs are placed facing each other at the centre of the room. One member from each team (the current speaker) occupies the centre chair at all times. The rest of their team sits in a semi-arc *directly behind* their speaker, facing the opposing pair. The "surrounded" geometry is thus two opposing benches facing each other across the two centre chairs.

The debate is conversational, not a sequence of speeches. Speakers address each other directly. There is no pre-set speaking order after the opening; the exchange follows the argument.

## *Chair Exchange Protocol*

Two mechanisms move speakers in and out of the chairs.

**Voluntary Pass**

### Core Rules for Debators

- **Stay in role.** You have been assigned a position; argue it, even if you disagree personally. The purpose is to stress-test positions, not to locate the comfortable middle. Thus, you may want to prepare for both sides of the argument in advance.
- **Handshake, then speak.** Every chair entry begins with a handshake. The ritual is not optional. It signals that the argument is about ideas, not the person.
- **Attack the argument, not the speaker.** Challenge the reasoning; do not make it personal.
- **Cite the text or case.** If you reference the assigned text or the session stimulus, name it and state what it shows. Vague gestures at "the story" are not arguments.
- **Concede precisely, then pivot.** If the opponent has made a point you cannot rebut, say so clearly and explain why your position survives it. A clean concession with a pivot is a stronger move than evasion.
- **Recognise quality across the line.** Knock your knuckles on the chair when the opposing or your speaker earns it. A debate that acknowledges good arguments is a better debate.
- **Use the pass.** Stepping down voluntarily when you are out of arguments is better than stalling. It gives your team a fresh voice. It is not a defeat.
- **Red-flag sparingly.** A red flag is a public vote of no-confidence in your current speaker. Reserve it for when the argument is being lost, not when you disagree with the phrasing. Remember: a red-flagged speaker may not return until every other team member has held the chair.
- **Team: confer quietly.** Members not in the chair may confer quietly with each other but may not coach their speaker audibly. Take notes; prepare to take the chair.

A speaker may tag themselves out at any point by saying *"I pass."* No team vote is required. The next available team member takes the chair. The outgoing and incoming speakers **shake hands** at the chair. *No handshake with the opponent;* this is a team substitution, not a challenge. A voluntary pass does *not* count against the rotation requirement: the passing speaker may return as soon as a teammate is willing to swap back.

**Red Flag (forced replacement)**
A team may replace their current speaker after a **minimum seat time of 60 seconds**. To red-flag:

1. A simple majority of the team members *not in the chair* raise a visible signal simultaneously (a card, a raised hand; agree on the signal before the session starts).

2. The moderator confirms the majority and calls the exchange.

3. The outgoing and incoming speakers **shake hands** at the chair. The incoming speaker then **shakes hands with the opponent** before speaking.

4. **Rotation rule:** a red-flagged speaker may not return to the chair until every other team member has held it at least once. Full rotation restores eligibility.

## *Debate Quality Mechanisms*

Two additional mechanisms are available to all participants. They are not used to change the chair; they are used to regulate the *quality* of the exchange, keep it constructive, and make intellectual honesty visible.

**Knocking (cross-team acknowledgment)**
When members of either team believe a speaker (including the opponent) has made an exceptionally strong argument, they acknowledge it by **knocking their knuckles on the armrest or seat of their chair**. The knock is the academic equivalent of applause: brief, audible, unmistakable.

1. Any number of participants may knock spontaneously; no majority is required. The knock is individual, not coordinated.

2. The moderator notes the moment but does not interrupt play. Knocking is logged informally as a marker of quality.

3. Play continues immediately. No chair change occurs. Knocking carries no mechanical penalty or reward; it is a public, embodied signal of respect for a well-made point.

*Purpose:* Builds a shared culture of recognising good argument. The physical gesture is harder to fake than raising a card and creates an immediate, visceral signal that something important was said.

**Point of Concession (speaker-initiated)**
A speaker in the chair may formally acknowledge that the opponent has made a point they cannot fully rebut. This is not a sign of weakness; it is the highest-quality move in academic debate.

1. The speaker raises an open hand (or a white card) and says: *"I concede: [state the opponent's point clearly and precisely]."* Vague concessions are not valid.
2. The speaker must then pivot: *"... my position survives because [explanation]."* A concession without a pivot is a capitulation and results in a speaker swap; the pivot is mandatory.
3. The moderator logs the concession and announces it briefly: *"Concession noted."*
4. The conceding speaker is **protected from a red flag for 30 seconds** after the concession. This makes intellectual honesty safe to demonstrate.
5. The opponent *may not repeat or pile on the conceded point* for the remainder of that speaking turn. The concession closes that line of argument; the debate must move forward.

*Purpose:* Makes weaknesses transparent rather than papered over. Rewards calibrated, honest reasoning. Keeps the debate from cycling on already-settled sub-questions.

## *Moderator Opening Script*

*This is a guide, not a script to read verbatim. Adapt it to the room.*

**1. Convene:**
The motion before us today is: [read the motion]. I am the moderator. I will manage chair exchanges, log concessions, and keep time. Please take your team seats.
**2. Assign teams and initial chairs:**
Team assignments are as follows (if possible done randomly): [read names and roles]. The Proposition team is [names]; the Opposition

team is [names]. The Proposition side opens with [Speaker A]; the Opposition side opens with [Speaker B]. Take two minutes now to review your opening position and agree on your red-flag signal.

**3. Read the stimulus and open:**

Before we begin, I will read the session stimulus. [Read the stimulus from the Session Stimulus section.]

[Speaker A] and [Speaker B]: please take the chairs, face each other, and shake hands. There are no opening statements; begin directly. The debate is open.

### *Moderator Closing Script*

The exchange is closed. I will now summarise the main fault lines. [Give a two-sentence neutral summary, noting any concessions that shifted the ground of the argument.] I hand over to the instructor for the wrap-up.

# *Bibliography*


N. Epley, A. Waytz, and J. T. Cacioppo. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4): 864–886, 2007. DOI: 10.1037/0033-295X.114.4.864.

European Parliament and Council of the European Union. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (ai act). Official Journal of the European Union, L series, 2024. Entered into force 1 August 2024.

L. Floridi and J. Cowls. A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 2019. DOI: 10.1162/99608f92.8cd550d1.

E. T. A. Hoffmann. *Der Sandmann*. 1816. Erstveröffentlichung in: *Nachtstücke*, Reimer, Berlin, 1817; Reclam Universalbibliothek Nr. 230.

F. Kafka. *In der Strafkolonie*. Kurt Wolff Verlag, Leipzig, 1919. Reclam Universalbibliothek Nr. 9900.

M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. DOI: 10.1109/MRA.2012.2192811. Translation of the 1970 original essay.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236): 433–460, 1950. DOI: 10.1093/mind/LIX.236.433.

K. Čapek. *R.U.R. (Rossum's Universal Robots)*. 1920. Uraufführung Prag 1921; Reclam Universalbibliothek Nr. 18418.

J. W. von Goethe. *Der Zauberlehrling*. 1797. Erstveröffentlichung in: *Musenalmanach*, 1798; in *Balladen*, Reclam Universalbibliothek Nr. 6845.

H. von Kleist. *Über das Marionettentheater*. 1810. Erstveröffentlichung in: *Berliner Abendblätter*, Dez. 1810; Reclam Universalbibliothek Nr. 9905.