Agency

2026-05-06 · cheerful mango Haubentaucher

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)*[1] by Karel Čapek (1920; Reclam UB 18418) *and Der Zauberlehrling*[2] 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?

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.*

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.

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 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.*

Asimov's Three Laws. (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 by humans, except where this conflicts with Law 1. (3) A robot must protect its own existence, except where this conflicts with Laws 1 or 2. Asimov spent his career writing stories about how these laws fail: conflicts, misinterpretation, and edge cases that no ranked rule set can anticipate.

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.