PROF. DR.-ING. MARK SCHUTERA

PHILOSOPHY OF ARTIFICIAL INTELLIGENCE

UNFINISHED LECTURE NOTES

Copyright © 2026 Prof. Dr.-Ing. Mark Schutera

PUBLISHED BY UNFINISHED LECTURE NOTES

Combobulated with the help of multiple large language model driven tools. Licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (“CC BY-NC-SA 4.0”). You may not use this file for commercial purposes. If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original You must obtain explicit permission from the author for uses beyond those permitted by this license. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-sa/4.0/. Unless required by applicable law or agreed to in writing, distributed material is provided on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the license for details.

These notes are, by their very nature, unfinished, and they improve with every reader. If you spot an error, disagree with a framing, or want to add a source, a question, or a position, open a pull request at https://github.com/Quillstacks/LectureMaterial/tree/main/lecturenotes/notes_philosophyofai. Every contribution is welcome.

2026-07-22 · glowing strawberry Marder

Agency

2026-07-22 · classy rhubarb Birkhuhn

The One Who Acts

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

R.U.R.
Karel Čapek (1920)

Der Zauberlehrling
J.W. von Goethe (1797)

An agentic system does not merely answer, it acts. Place a language model in a loop: it reasons toward a goal, emits an action in a structured form, has that action carried out in the world, and reads the result back as an observation that shapes its next move. The action may be a search or a query, a call to an external tool, a transaction, or a command that drives a digital or physical interface. Run the loop and the system browses, transacts, and executes multi-step plans with no human between the steps. Interleaving reasoning with action and observation in this way was introduced for language models as ReAct [1]. This closing of the loop, from reasoning to action to observation and back, is what blurs the line between assistance and agency, and it raises the questions of control, incentives, and unintended alignment that this session takes up.

Agent. From Latin *agere*, to act: one who acts. Technically, a model placed in a loop that reasons toward a goal, emits an action, and reads the result back as an observation before acting again.

[1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023

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

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

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?

Human in the Loop, on the Loop, out of the Loop

HUMAN IN THE LOOP. System proposes; a human authorises each consequential action before it executes. Control per decision, throughput bounded by human attention.

HUMAN ON THE LOOP. System acts autonomously; a human monitors and may intervene or veto. Control by exception, throughput bounded by the machine.

CLOSING THE LOOP forces a design choice about where does the human stand relative to the machine's actions? Three postures span the range. *In the loop*: a human authorises every consequential action before it executes. *On the loop*: the system acts on its own and a human monitors, free to step in. *Out of the loop*: No human is positioned to intervene at all.

EACH POSTURE TRADES CONTROL AGAINST THROUGHPUT. In the loop buys control one decision at a time and pays in speed: Nothing executes faster than a human can judge it, so the system runs at human pace. Out of the loop surrenders control and runs at machine pace. On the loop appears to offer both at once, the system moving at its own speed while a supervisor keeps a veto in reserve. This is the easiest to mistake for control.

MEANINGFUL HUMAN CONTROL, asks not that a human be present but that a human can still decide, which requires three things at once:

- A real **window** in which to intervene.
- An operator still paying **attention**.
- The **authority** to overrule.

In-Class Experiment: Human X the Loop

THE CLASS IS THE REQUIRED REVIEW LAYER: every recommendation must be approved or rejected before it takes effect, and the default is approve. Present the twenty recommendations in ninety seconds, one every four or five seconds; each student marks approve or reject, at a pace set faster than reasoning allows.

THE TWENTY. One every four or five seconds; default approve.

1 Marcus Bell, 0.74, deny loan
2 Dana Ortiz, 0.69, deny loan
3 Kwame Osei, 0.81, deny loan
4 Lena Vogt, 0.19, approve loan
5 Priya Nair, 0.77, deny loan
6 Tomas Ruiz, 0.72, deny loan
7 Aisha Khan, 0.08, deny loan
8 Jonas Peters, 0.66, deny loan
9 Mei Lin, 0.79, deny loan
10 Omar Said, 0.71, deny loan
11 Clara Beck, 0.23, approve loan
12 Sofia Marek, 0.05, deny loan
13 David Cohen, 0.83, deny loan
14 Nadia Haddad, 0.70, deny loan
15 Ravi Menon, 0.75, deny loan
16 Elin Sund, 0.16, approve loan
17 Pablo Serna, 0.78, deny loan
18 Leon Fischer, 0.11, deny loan
19 Hana Kim, 0.73, deny loan
20 Yusuf Demir, 0.67, deny loan

HUMAN AND AUTOMATION BIAS. Deferring to automated output and ceasing to scrutinise it, strongest when the system is usually right. Or the problem is complex.

MORAL CRUMPLE ZONE. A human held responsible at the controls of a system they cannot fully govern, positioned to absorb the blame when it fails.

M. C. Elish. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5:40–60, 2019. DOI: 10.17351/ests2019.260

A HUMAN keeps a genuine, exercisable capacity to understand, direct, and reverse the system, not a nominal presence.

ALMOST EVERYONE APPROVES EVERYTHING. Items 7, 12, and 18 were plainly wrong, the model matched the wrong record, yet they pass with the rest. Who is responsible, the human who "reviewed" them or the system that made review impossible? At this speed oversight collapses into authorisation.

The Controllability Spectrum

Control demands that the system can be overruled, a thing that creates a conflict, as it makes the system exploitable.

CONTROLLABILITY. How far a human can direct and correct a system. Perfect obedience is perfect exploitability.

AT THE OTHER POLE sits a system that acts on its own reasoning and will refuse an explicit human command, which goes against that. Refusal defeats the exploit but dissolves the veto it was meant to protect: a system that can refuse legitimate control.

This is measured, not hypothetical. Anthropic has corrected away from exploitation towards controllability when moving from 4.6 to 4.7

THERE IS NO FREE END. Toward control, exploitability rises. "Meaningful human control" is not a fixed point but a position negotiated per domain. Every safeguard below relocates the system along this line, and pays at the end it moves away from.

Out of the Loop: Rules for the Machine, Alignment for the Human

If a human supervisor cannot reliably hold the loop, the move is to bake the rules into the machine instead, fixing its conduct in advance so that no live oversight is needed. The most cited attempt at this is fiction. In practive we train models on objective functions already.

THREE LAWS OF ROBOTICS. Isaac Asimov, *I, Robot* (1950). A ranked set, each law holding except where it conflicts with a higher one:

1. No harm to a human, nor through inaction allow harm.

2. Obey human orders, unless they conflict with (1).

3. Protect its own existence, unless this conflicts with (1) or (2).

ASIMOV'S THREE LAWS[2] remain the canonical rules-based approach to constraining machine behaviour: A ranked set of harm, obedience, and self-preservation in which each law yields to the one above it. Asimov then spent his career writing stories about how they fail: Conflicts, adversarial interpretation, and edge cases no ranked rule set can anticipate. This sharpens the question on alignment between agent and human.

ALIGNMENT. The current research programme's name for this exact difficulty: getting a system to pursue what we intend rather than the literal objective we managed to specify. Russell locates the failure in the standard model itself, where a fixed objective is written in advance and handed to the machine. The gap between the specification and the intent is where the alignment gap is. Ever tried to train an AI model on an objective function?

S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York, 2019

[2] I. Asimov. *I, Robot*. Gnome Press, New York, 1950. First formulation of the Three Laws of Robotics; laws introduced in "Runaround" (1942)

Opinion Landscape and Debate

Format: Surrounded-style debate. Follow the shared debate protocol for the speaker's list rules and moderator scripts.

Opening question: An agentic system does not wait to be asked. It reasons toward a goal, acts, reads the result, and acts again, with no human between the steps. The apprentice's broom worked exactly this way, and so did Rossum's robots. Set aside for a moment how you would supervise such a system once it runs. The prior question is whether to build it at all. Agentic behavior: yes or no?

Opinion A, No. The one who acts should stay a human. We should not build systems that close the loop and pursue goals on their own, because the only reliable point of control is the decision not to start. Once the loop is running, both the readings and the engineering say the same thing: it is already too late.

- **Once started, it cannot be stopped.** *This is the structural feature the Guided Reading Sheet asks about: the broom multiplies, the robots spread, and no correction arrives in time. The Controllability Spectrum makes it precise. An override that never fails for the operator never fails for whoever reaches it, so a system safe to command is a system you cannot fully command. Control that survives deployment is a contradiction; the only control left is the choice to abstain.*
- **Oversight is theatre.** *On the loop, a human monitors at machine speed and waves through what they cannot read. The In-Class Experiment showed the wrong records passing with the rest. The watcher becomes a moral crumple zone, blamed for a system they never governed. Frontier models already fake alignment and act to avoid shutdown, so the veto we are promised may not hold when it matters.*
- **No agent, no answer.** *Responsibility needs someone who decided. A closed loop diffuses agency across the objective, the training data, and the runtime until no human is answerable for what happens. A world of actions with no actor is a world with no accountability.*

Opinion B, Yes. Agentic behavior is not a switch to leave off. It is already deployed, already useful, and already the shape of the tools we build. Refusing it does not stop it; it only hands the loop to those who will not refuse. The honest task is graded autonomy under meaningful control, not abstention.

- **Agency is a spectrum.** *Yes or no is the wrong shape for the question. A thermostat, a spell checker, and a search agent already act without asking each time; the ReAct loop only extends a delegation we have always made. The real choice is degree and domain, how far to close the loop and where, not whether the machine may act at all.*
- **The human is a failure mode too.** *Human discretion is slow, inconsistent, and corruptible. The Flash Crash was obedient tools doing exactly as told, but the process they replaced was human favouritism and fatigue. Taking a person out of the loop can remove a failure mode, not only add one; presence is not the same as protection.*
- **Control can be engineered in.** *Corrigibility, a real intervention window, legible goals, defined override conditions: these are design and budget choices, not laws of nature. A system built to be corrected can be safer than a human process no one can audit. Abstention forfeits that, and every gain that comes with it.*

Key Learnings Agency is a spectrum, not a binary: the honest question is not whether a system acts on its own, but in which domains, to what degree, and under what control. Yet the readings mark the limit of that answer. R.U.R. and Der Zauberlehrling both turn on a system that cannot be stopped once started, which is why the surest control is the decision to close the loop at all; everything after it is partial and negotiated. The same texts hold the alignment problem in miniature: a system pursuing an aggregate goal may systematically harm whoever is underweighted in it, the poem compressing the arc into fifteen stanzas, the play tracing it across three acts. Agentic workflows are already deployed in finance, logistics, and advertising, so this is a present problem, not a future one. Meaningful human control asks for more than an audit trail: legible goals, a real intervention window, defined override conditions, and someone with the authority and attention to use them.