ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS

Lecture 10a – Frontier AI, Governance, and Social Impact

Chizhi Chris ZHANG

zhangchizhi@ciomp.ac.cn
Advanced Computing and Digital Technology Research Center
University of Chinese Academy of Sciences

Spring 2026

Today's Question

What we are trying to answer

As AI systems become more capable, how should society use them without losing trust, fairness, and human agency?

Why this final lecture matters

Earlier lectures mostly asked what AI can do. The final lecture asks a different question: how should we live with systems that can increasingly affect knowledge work, services, and institutions?

What changes from AI9

AI9 focused on agents and multi-step action. AI10 widens the lens from system design to social design: responsibility, governance, institutions, and human judgment.

From AI9 to AI10

Last time

We asked how agents plan, use tools, remember, and act across several steps.

Today

We ask what happens when those capabilities move from labs and demos into education, work, public services, and everyday life.

One sentence

AI9 was about how agents work. AI10 is about what their spread means.

A Historical Anchor

Long arc

The early public question was whether a machine could hold a convincing conversation.

Current reality

Now the frontier question is broader. We care about conversation, tool use, task completion, and the real consequences of machine decisions.

Why this belongs in the final lecture

The public conversation changed from “Can it imitate?” to “Can it be trusted in the real world?”

What Frontier AI Means Today

Current capability trends

- stronger reasoning
- broader multimodal input
- longer context handling
- more tool-connected workflows

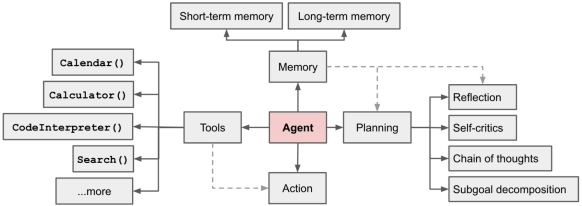
Boundary reminder

High capability does not mean universal reliability. The systems are more useful than before, but still uneven, fragile, and context-dependent.

A realistic reading

The same model may draft a strong report, fail on a spreadsheet detail, and then recover if a human reframes the task well.

Why the Frontier Now Looks Agentic



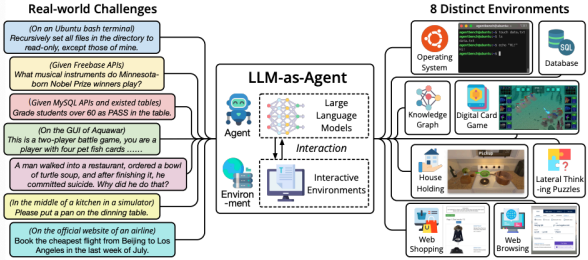
What changed in public perception

The frontier no longer looks like a model that only talks. It increasingly looks like a system that can coordinate tools, memory, and workflows.

Why people notice this quickly

A chatbot feels like conversation. An agent that reads email, checks a calendar, drafts a reply, and asks for approval feels like a coworker with limits.

How to Read Frontier Benchmarks



What benchmarks are good for

They are useful signals about progress.

What they are not

They are not guarantees that a system will behave well in a messy real environment with unclear instructions, human stakes, and institutional rules.

Why Multi-Step Capability Raises the Stakes

The social shift

Once systems are planning multi-step work rather than only generating text, the question shifts from content quality alone to responsibility for actions and outcomes.

What changes in practice

A wrong paragraph is annoying. A wrong chain of search, routing, recommendation, and action can quietly affect real people before anyone notices.

Why governance enters here

The more autonomous the workflow becomes, the more important it is to decide who can authorize, monitor, and interrupt the system.

Where the Real Value Appears

Where value usually appears first

- drafting, summarization, comparison, and synthesis
- translation, navigation, and first-line service response
- documentation, debugging support, and workflow assistance

What these cases have in common

They involve a lot of language, repeated patterns, and work that benefits from a faster first draft before a human makes the final call.

Capability Only Matters Inside a System

The point

Value does not come from model quality alone. It comes from the whole stack: interface, workflow design, permissions, monitoring, and the human process around the model.

Why this is easy to forget

People often compare models as if the best model automatically creates the best social outcome. In reality, the surrounding workflow often matters just as much.

The same model can feel helpful in one setting and unacceptable in another because the rules, oversight, and incentives around it are different.

The Same Model Can Feel Very Different

A simple comparison

The same base model may be welcomed as a study assistant, questioned as a hiring screener, and rejected as an automatic decision maker in a benefits office.

Why the reaction changes

The stakes, evidence standard, appeal rights, and social expectations are different even when the underlying capability looks similar.

The lecture point

That is why governance cannot be copied from one domain to another as if every use case carried the same moral weight.

From Model Quality to User Value

Model level

Accuracy, fluency, retrieval quality, and grounded reasoning.

System level

Latency, reliability, permissions, escalation design, and a workflow users can actually trust.

This is why governance is not an “extra” after the technical work. It is part of turning capability into safe value.

A Public Service Scenario

Problem

A city office receives a huge number of repetitive citizen questions in different languages and formats.

What AI can contribute

Round-the-clock first-line support, routing, form guidance, and clearer explanations of standard procedures.

What still must stay human-led

Exceptions, policy interpretation, appeals, and accountability.

A Campus Scenario

A realistic use case

A university support bot helps students interpret course rules, find office procedures, compare program requirements, and prepare before they talk to a real staff member.

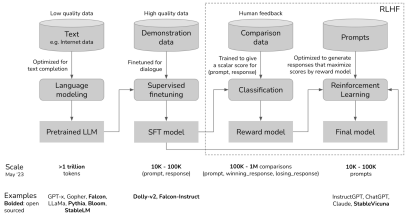
Why students usually like this

It saves time on repetitive questions and lowers the barrier for students who feel unsure where to start.

Where the boundary still matters

Once the case becomes personal, exceptional, or high-stakes, the answer cannot stop at a smooth paragraph. Someone accountable has to step in.

Why Alignment Pipelines Matter



Why this belongs in AI10

Powerful systems are not useful only because they predict text well. They also go through post-training choices that try to shape helpfulness, refusal behavior, tone, and policy compliance.

Why ordinary users should care

Two systems with similar raw capability can feel very different because one is better post-trained to ask clarifying questions and stay within policy boundaries.

The Impact on Work Is Uneven

Likely early impact

Routine cognitive work with repeatable formats, clear templates, and lots of documents.

Likely resilient areas

Roles that require trust, negotiation, accountability, physical context, or high-stakes judgment under uncertainty.

The question is not simply “Which jobs disappear?” It is also “Which parts of jobs change first?”

Education Changes Too

Opportunity

Personalized feedback, tutoring support, alternative explanations, and practice generation at scale.

Risk

If assessment design does not adapt, students may outsource too much thinking and lose some independence.

What educators now have to redesign

Not only content, but also how we evaluate understanding, process, and responsibility.

Healthcare and Public Interest

Positive potential

Documentation support, triage support, search over guidelines, and clearer communication with patients or citizens.

Governance requirement

Clinical or legal decisions need strict oversight, auditability, and human sign-off.

The higher the stakes, the less acceptable it is to hide behind fluent output.

The Digital Divide Question

The concern

Benefits may concentrate where compute, data, AI literacy, and institutional resources are already strong.

Why policy enters the story

Access, training, and shared educational resources matter if AI capability is not supposed to widen existing inequality.

Technology spreads unevenly unless institutions actively design for broader access.

What Society Is Really Debating

Not only “Can AI do this?”

The harder question is “Who decides how AI should be used in this context, and who is responsible when it goes wrong?”

Core tension

Innovation speed, public trust, legitimacy, and safety do not automatically move together.

That tension is why AI governance is a live social issue, not just a technical footnote.

When Responsibility Gets Blurry

A common failure pattern

The model gives a polished answer, a staff member trusts it too quickly, the user follows it, and the damage appears only later.

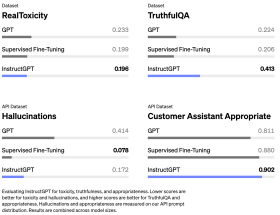
Why this matters

When responsibility is unclear, every actor is tempted to say the failure came from somewhere else: the model, the interface, the policy, or the user.

The governance question

Good governance makes roles explicit: who built the model, who deployed it, who approved the workflow, and who can intervene when the system is wrong.

Why Hallucination Still Matters



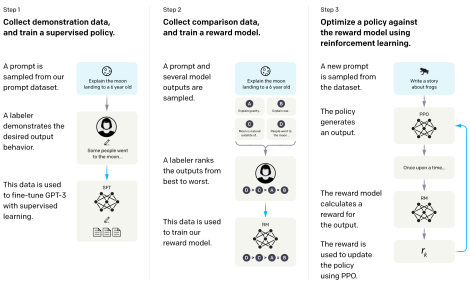
Why this matters outside the lab

Confidently wrong output can mislead students, patients, officials, and ordinary users even when the system sounds polished and helpful.

A familiar failure

When a model invents a citation or confidently misstates a rule, the polished style can discourage the user from double-checking.

Alignment Is an Ongoing Process



The practical lesson

Alignment is not one switch you flip once. It is a continuing cycle of data choices, policy updates, monitoring, and correction after deployment.

Why the Objective Is Hard

$$\begin{aligned}\text{objective}\,(\phi) =& E_{(x,y)\sim D_{\pi_\phi^{\mathrm{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\mathrm{RL}}(y \mid x)/\pi^{\mathrm{SFT}}(y \mid x)\right)\right] + \\ & \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\mathrm{RL}}(x))\right]\end{aligned} \tag{2}$$

What keeps colliding

Truthfulness, helpfulness, harmlessness, user preference, and business goals do not always point in the same direction.

Governance Has Layers

Model layer

Training choices, safety testing, and robustness checks.

System layer

Access control, logs, escalation paths, and fallback behavior.

Institution layer

Law, standards, auditing, appeal mechanisms, and accountability.

All three layers matter because a system can be technically strong and still be socially unacceptable if the workflow around it is careless.

Why Trust Is Hard to Rebuild

Trust usually breaks faster than it grows

One visible failure in a sensitive setting can shape public opinion more strongly than many quiet successes.

What helps

Clear disclosure, good fallback behavior, visible human accountability, and a real path for appeal matter as much as raw model quality.

That is why governance is not only about restriction. It is also about making adoption legitimate enough that people are willing to use the system in the first place.

A Practical Risk Checklist

Before deployment

- define forbidden actions
- design a human escalation path
- test hard prompts and edge cases
- decide how incidents will be logged

After deployment

Monitor drift, review incidents, update policy, and keep the system aligned with how it is actually being used in the real world.

What Skills Grow in Importance

Human strengths that matter more

- framing ambiguous problems
- making ethical tradeoffs
- deciding under uncertainty
- taking responsibility

A better way to think about AI

Use it as a reasoning partner and productivity amplifier, not as a substitute for judgment.

How Students Should Use Frontier AI

Good habits

- ask for assumptions explicitly
- compare multiple perspectives
- verify important claims with evidence
- treat outputs as drafts when stakes are high

Bad habits

Copy without checking, ignore sources, or treat fluency as proof.

Career Strategy for Non-AI Majors

Do not chase every model release

Stable value still comes from a strong domain foundation.

Then add AI leverage

Use AI to improve speed, breadth, explanation quality, and the range of options you can consider in your own field.

The goal is not to become a copy of the model. The goal is to become better at your own discipline with AI as leverage.

Why AI10 Leads to NN10

AI10 question

How should society govern and use frontier AI?

NN10 question

If we want useful and governed systems, how do we actually make them efficient, observable, aligned, and deployable at scale?

The social question naturally points to the systems question.

Summary

- Frontier AI changes the conversation because capability is increasingly connected to multi-step workflows and real institutions.
- The value of AI depends on the whole system around the model, not only on raw model quality.
- Work, education, health, and public service will all be affected, but not in the same way or at the same speed.
- Alignment and governance are ongoing processes because powerful systems can still be wrong, biased, or misused.
- Human judgment matters more, not less, when systems become stronger and easier to rely on.

After the Course

What should stay with you

When you encounter a new AI system, do not stop at asking whether it looks impressive. Ask what it is allowed to do, where it can fail, who is responsible, and how people can intervene.

A useful habit

Carry the full chain together: model capability, system design, human oversight, and social consequence. That is the difference between watching AI news and actually understanding AI.

Thank You