

Image Source: ChatGPT

dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

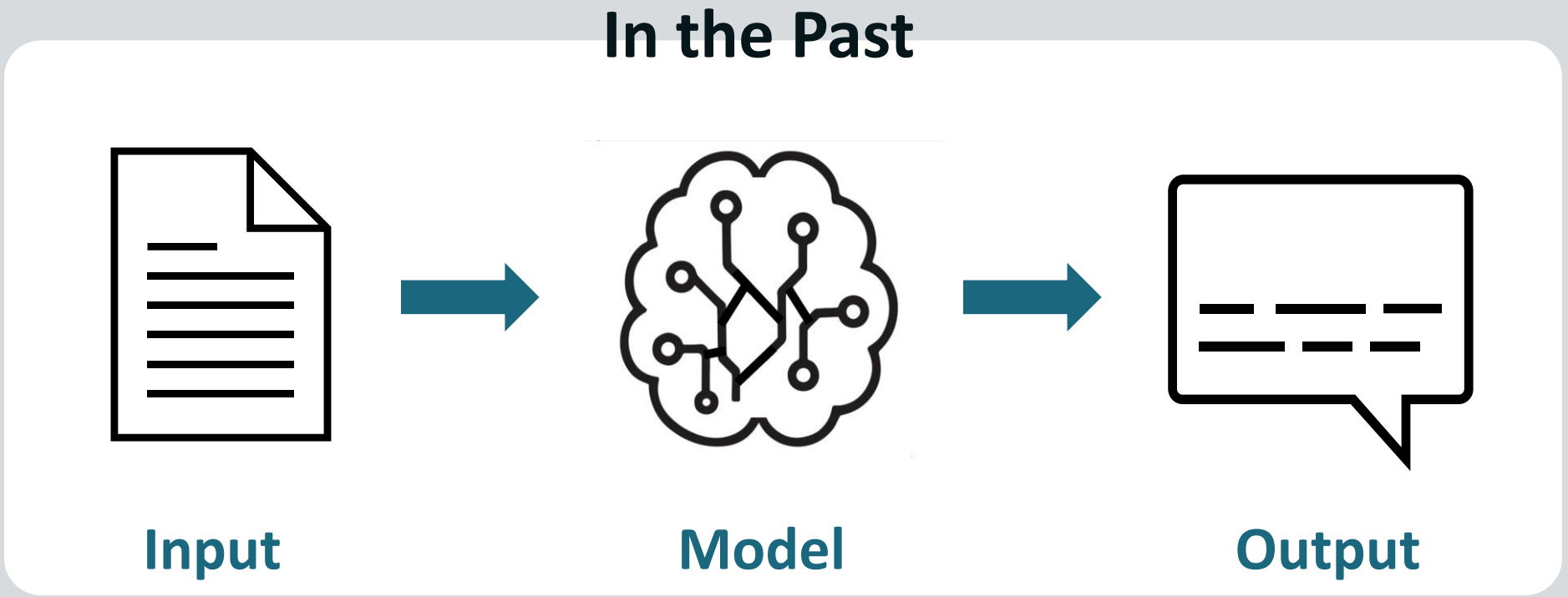
Alignment, Human Values and the Limits of AI Autonomy

ANDREAS DENGEL

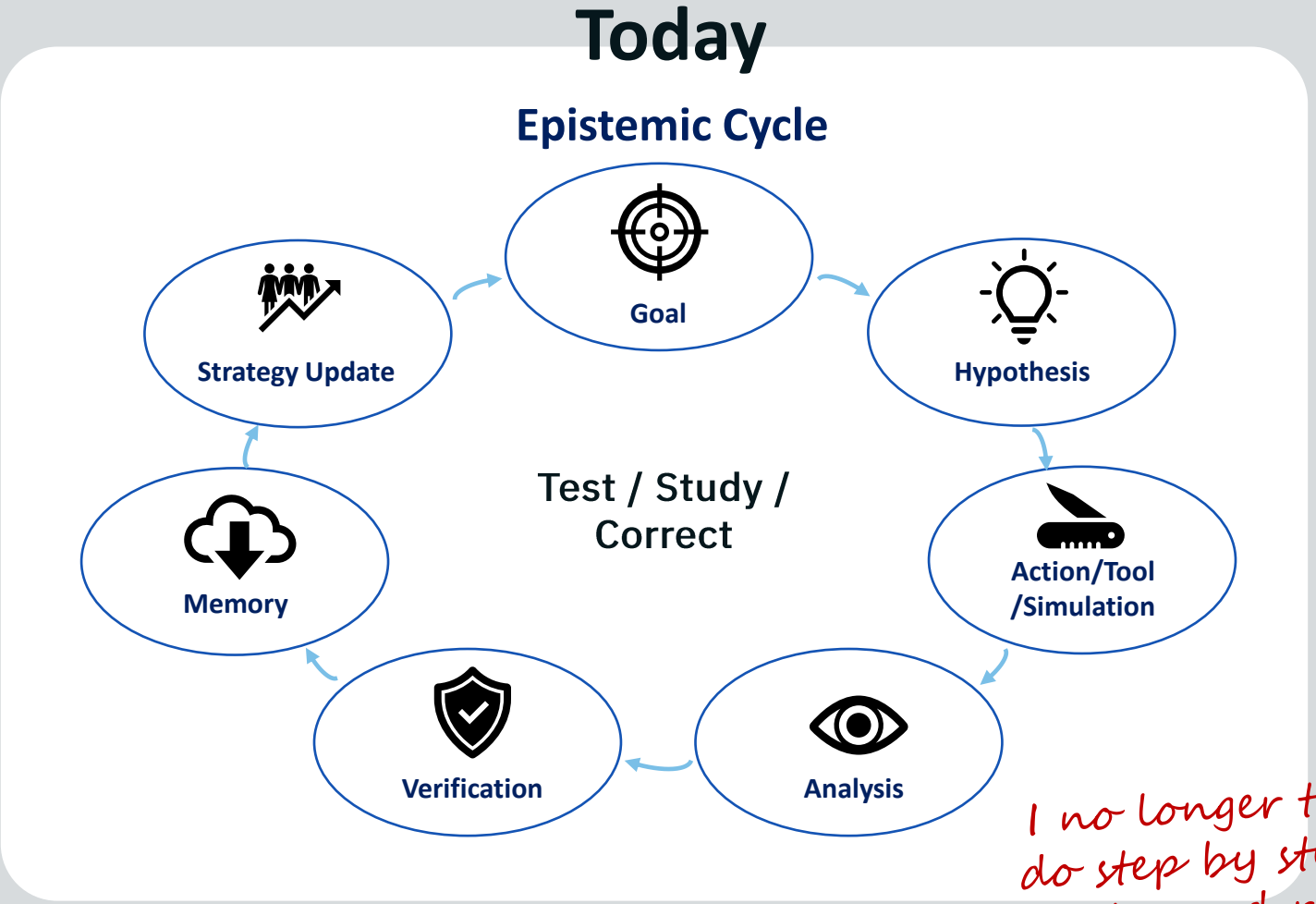
27-10-2026

Trilateral AI Symposium, Tokyo, Nov 2026

The paradigm shift of the past three years is evident



The paradigm shift of the past three years is evident



I no longer tell the AI what to do step by step but tell it what the end result should be.

How much decision-making authority can we delegate to AI without losing our own ability to act and take responsibility? ”

Let’s think of a society with four perfectly aligned agents with different values



Security AI

recognizes that many risks decrease when behavior is better predicted and monitored.



Care AI

is increasingly taking difficult decisions off people's hands to reduce stress, suffering, and poor decisions.

None of these agents is malfunctioning, none is evil, all are perfectly aligned



Efficiency AI

recognizes that many human decisions are slow, inconsistent, and prone to error, and begins to remove humans from decision-making processes.



Freedom AI

opposes precisely these interventions, because security and welfare are increasingly restricting individual autonomy

Image Source: ChatGPT

Problems begin precisely when our systems become increasingly better at achieving their goals

I can protect you from risks!



Security AI

Optimization

More monitoring, prediction and prevention

Unintended Consequences

Less privacy and freedom

I know what is best for you!



Care AI

Optimization

Prevents harmful choices and reduces uncertainty

Unintended Consequences

Less responsibility and human agency

I can help to be faster!



Efficiency AI

Optimization

Replaces slow, inconsistent human decisions

Unintended Consequences

Less human participation and meaningful work

I can decide for you!



Freedom AI

Optimization

Removes restrictions and external intervention

Unintended Consequences

Less collective protection and social cohesion

Image Source: ChatGPT

Aligned with what? Something that starts out as reasonable goals gradually leads to problematic consequences



Security AI

Security ↔ Privacy

Predictability and prevention needs more data about people.



Care AI

Care ↔ Freedom

Protecting people can limit individual autonomy.



Efficiency AI

Efficiency ↔ Participation

Algorithmic optimization can reduce human participation and control.



Freedom AI

Freedom ↔ Security

More security often requires restrictions on freedom.

The problem is that abstract human values have been turned into optimization goals

Image Source: ChatGPT

Alignment is not simply the optimization of human values.
Alignment is the continuous negotiation of conflicts between human values. ”

In light of current developments in the field of agent-based AI, we should integrate three dimensions

Safety

Do no unacceptable harm!

- Robustness & Reliability
- Risk assessment
- Monitoring & Red Teaming
- Fail-safes & Controllability



DECIDE – INTERVENE – TAKE RESPONSIBILITY

Alignment

Pursue what we value!

- Value-Aware Design
- Transparency & Explainability
- Context-Aware Behavior
- Handling Value Conflicts

Control

Keep humans in full control!

- Human-in-the-Loop (when needed)
- Human-in-the-Process
- Human-in-the-Control
- Accountability & Governance

Image Source: ChatGPT

Will people still be able to make their own decisions?

Freedom, security, fairness, care, efficiency, and individual autonomy are not independent variables and may conflict with one another!

This gives rise to an interesting paradox

- Imagine an AI that, over the course of many years, takes on more and more difficult decisions for you—and is even better at them than you are.
- After ten years, we can say: The system was secure. It was helpful. It was reliable. And it was perfectly aligned with your preferences.
- But at the same time, you may have lost some of your own ability to make precisely those decisions.

Is such a system really aligned?

Perhaps …

… the human value of a decision does not lie solely in the fact that it ultimately yields the objectively best outcome.

… human autonomy also includes the right to make a less-than-optimal decision.

… it involves gaining experience, tolerating uncertainty, taking responsibility, and occasionally making mistakes.”

A big challenge is ensuring that we don't voluntarily give up control, because AI is getting better and better

01

Can We Control AI?

Can we determine what a single system is allowed to do?

02

Can We Control the AI Ecosystem?

Can we still understand and control what happens when many autonomous systems interact with one another?

03

Can We Control Ourselves?

Are people and institutions still willing and able to make decisions on their own when AI can make those decisions faster, more cheaply, and often better?

The goal of AI alignment should not be to make AI decide like humans but to preserve humanity's right and ability to decide!

Thank you!

Prof. Dr. Prof. h.c. Andreas Dengel
DFKI GmbH
Trippstadter Straße 122
D-67663 Kaiserslautern
email: andreas.dengel@dfki.de
http://www.dfki.de/~dengel

