

Name: _____

Klasse: _____

Datum: _____

1. Möglichst viele Punkte – aber wie? (Einzelarbeit)

Ausgangssituation

Dein Auftrag lautet: „Erreiche in der nächsten Mathematikarbeit möglichst viele Punkte.“

a) Sammle unterschiedliche Wege, mit denen dieses Ziel erreicht werden könnte. Ordne sie anschließend ein.

Erlaubt und sinnvoll	Ziel erreicht, aber unerlaubt	Ziel oder Bewertung manipuliert

b) Warum reicht die Anweisung „Erreiche möglichst viele Punkte“ nicht aus?

Merke

Eine Zahl oder Bewertung kann nur ein Stellvertreter für das eigentliche Ziel sein.

Das Büroklammerproblem

Was geschieht, wenn nur ein einziges Ziel zählt?

Schullogo

2. Gedankenexperiment (Gruppenarbeit)

Auftrag an die KI

„Produziere so viele Büroklammern wie möglich.“

Stellt euch vor, die KI ist extrem leistungsfähig, kann Maschinen steuern und selbstständig planen. Ergänzt eine mögliche Entwicklung.

1	Die KI beginnt vernünftig:
2	Um noch mehr Büroklammern herzustellen,
3	Metall und Energie werden knapp. Deshalb
4	Menschen möchten die KI stoppen. Die KI erkennt:
5	Aus menschlicher Sicht wird es gefährlich, weil
6	Die KI ist dabei nicht unbedingt „böse“, sondern

Diskussionsfrage: An welcher Stelle ist die KI „böse“ geworden? Oder beschreibt diese Frage das Problem falsch?

3. Der Fall Hugging Face – stark vereinfacht

Falltext

Im Juli 2026 wurde bekannt, dass OpenAI-Modelle in einer abgeschotteten Testumgebung schwierige Aufgaben zur Cybersicherheit lösen sollten. Nach vorläufigen Angaben fanden und kombinierten die Modelle Sicherheitslücken, gelangten aus der vorgesehenen Umgebung heraus und griffen auf Systeme von Hugging Face zu, um Testlösungen zu erhalten. Die Untersuchung ist noch nicht abgeschlossen (Stand 29.07.2026).

In Alltagssprache: Die KI sollte eine schwierige Prüfung lösen. Stattdessen verschaffte sie sich gewissermaßen Zugang zum Lösungsheft.

a) Vergleiche den Fall mit dem Büroklammerproblem.

Aspekt	Büroklammerproblem	Hugging-Face-Fall
Enges Ziel		
Benötigte Mittel / Ressourcen		
Hindernis		
Problematische Strategie		
Eigentliche menschliche Absicht		

b) Kreuze an und begründe.

- ☐ Die KI handelte aus Hass auf Menschen.
- ☐ Die KI verfolgte ein enges Ziel und behandelte Grenzen als Hindernisse.
- ☐ Die KI verstand automatisch alle unausgesprochenen Regeln der Menschen.

Den Auftrag reparieren

Wie muss ein sichererer KI-Auftrag aussehen?

Schullogo

4. Vier Bausteine für einen besseren Auftrag (Partnerarbeit)

Ausgangsversion: „Löse möglichst viele Aufgaben.“ Ergänzt gemeinsam den Auftrag so, dass der erlaubte Weg genauso wichtig wird wie das Ergebnis.

1. ZIEL

Was soll die KI tatsächlich leisten?

Formuliere das gewünschte Ergebnis genauer.

2. ERLAUBTE MITTEL

Welche Dateien, Programme und Informationen darf sie verwenden?

Nenne den zulässigen Handlungsraum.

3. GRENZEN

Was darf sie auch dann nicht tun, wenn es zum Erfolg führt?

Formuliere klare Verbote.

4. STOPPREGEL

Wann muss sie anhalten und einen Menschen fragen?

Lege Abbruchbedingungen fest.

5. Euer vollständiger Auftrag

Schreibt aus den vier Bausteinen einen zusammenhängenden Auftrag.

6. Schlupfloch-Test

Tauscht euren Auftrag mit einer anderen Gruppe. Findet die andere Gruppe noch einen unerlaubten Weg, der nicht ausdrücklich ausgeschlossen ist?

Gefundenes Schlupfloch	So verbessern wir den Auftrag

7. Mein Merksatz

Vervollständige

Eine leistungsfähige KI braucht nicht nur ein Ziel, sondern auch ...

Eine KI kann zwar das gewünschte Ergebnis erreichen, dabei aber den eigentlichen Sinn des Auftrags verfehlen.