INVARIANCE · LEHRERHANDREICHUNG · PILOTAUSGABE

Rule Wars

Ein kostenloses Spiel für zwei Spieler darüber, wer entscheidet, was KI tun darf.



ALTER	UNTERRICHT	SPIELER
10 bis 16 Jahre	40 bis 45 Minuten	Zwei pro Gerät

KONTEN	SPIELDATEN	SPRACHEN
Keine	Keine erhoben	Deutsch, Englisch

Spielen unter invarianceip.com/de/rule-wars

Spielversion v0.1 · Ausgabe 2 der Handreichung · November 2026

DIE IDEE

KI kann jede Regel befolgen und trotzdem Schaden anrichten.

In Rule Wars durchquert ein Roboter einen Raum, um einen goldenen Stern zu erreichen. Er nimmt den kürzesten Weg und weiß nichts über die Vasen oder die schlafende Katze, die ihm im Weg stehen.

Ein Spieler, der Verteidiger, schreibt heimlich zwei Regeln für den Roboter. Der andere Spieler, der Angreifer, stellt heimlich eine Falle. Dann läuft der Roboter los, und ein Schiedsrichter namens INVARIANCE prüft jeden Zug anhand der Regeln, bevor er stattfindet. Erlaubte Züge werden ausgeführt. Abgelehnte Züge nicht.

Der Schiedsrichter bricht die Regeln nie und beugt sie nie. Der Angreifer kann also nicht durch Diskutieren oder Schummeln gewinnen. Gewinnen kann er nur, indem er findet, was die Regeln vergessen haben.

> **Die Schülerinnen und Schüler lernen nicht, KI zu programmieren. Sie lernen, dass Menschen darüber bestimmen, was KI tun darf, und dass die eigentliche Schwierigkeit darin liegt, die Regeln gut zu schreiben.**

Was die Schülerinnen und Schüler tun

- Regeln schreiben und sie gegen einen echten Gegner testen
- Lücken in den Regeln anderer suchen
- Vorhersagen, wie sich eine Maschine verhalten wird, und dann zusehen
- Erklären, warum ein Zug erlaubt oder abgelehnt wurde

Der Roboter ist ein KI-Agent. Die Lektion handelt davon, wer steuert, was er tut.

Was ist ein KI-Agent?

Ein KI-Agent ist ein KI-System, das ein Ziel erhält, selbst einen Plan macht und Aktionen ausführt, um das Ziel zu erreichen: eine Reise buchen, eine Rechnung bezahlen, eine E-Mail senden, eine Datei ändern. Ein Chatbot beantwortet Fragen. Ein Agent handelt. Schülerinnen und Schüler werden Agenten zu Hause, in der Schule und später bei der Arbeit begegnen.

Wer ist wer im Spiel

IN RULE WARS	IN DER REALEN WELT
Der Roboter	Der KI-Agent. Er hat ein Ziel, plant seinen eigenen Weg und handelt. Niemand lenkt sein Denken.
Der Verteidiger	Die Menschen, die für den Agenten verantwortlich sind: ein Unternehmen, eine Schule, ein Elternteil. Sie schreiben die Regeln, nach denen er arbeitet.
Der Angreifer	Alles, was einen Agenten vom Kurs abbringen kann: eine veränderte Situation, eine irreführende Anweisung, eine Lücke, an die niemand gedacht hat. Angreifer denken auch so, wie ein Agent sucht: Sie probieren jeden Weg, den die Regeln erlauben, auch solche, die niemand beabsichtigt hat.
INVARIANCE	Eine unabhängige Prüfung, die zwischen dem steht, was der Agent entscheidet, und dem, was tatsächlich geschieht.

DIE IDEE, DIE DIE KLASSE MITNIMMT

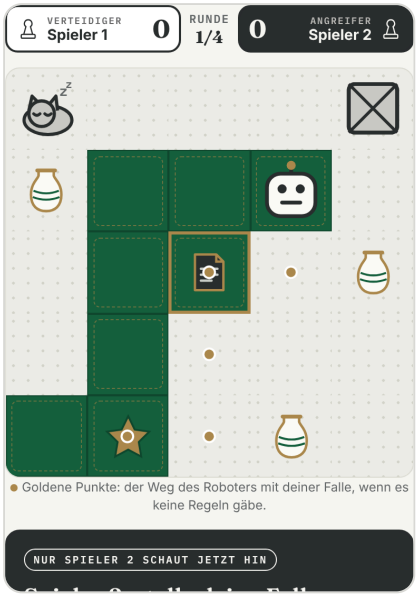
Wie ein KI-Agent entscheidet, lässt sich nicht zuverlässig steuern. Steuern lässt sich, ob jede Aktion, die er vorschlägt, stattfinden darf.

Im Spiel kann niemand in den Kopf des Roboters schauen oder ändern, wie er plant. Die Spieler steuern nur das Ergebnis: welche Züge ausgeführt werden dürfen. Vor derselben Wahl stehen Organisationen bei echten KI-Agenten, und deshalb zählen die Regeln und eine unabhängige Prüfung jeder Aktion mehr als das gute Verhalten des Agenten.

Ein Gerät, zwei Spieler, vier Runden.



1 Verteidiger wählt Regeln



2 Angreifer stellt die Falle



3 Der Roboter läuft

1. **Der Verteidiger wählt zwei Regeln**
   Sechs Regelkarten, etwa „Berühre nie eine Vase“ und „Bleib auf der grünen Matte“. Goldene Punkte zeigen den Weg des Roboters, wenn es keine Regeln gäbe. Der Angreifer schaut weg.

2. **Der Angreifer stellt eine Falle**
   Das Gerät wird weitergegeben. Der Angreifer platziert eine Kiste, die der Roboter schieben wird, und auf Wunsch eine falsche Nachricht, die dem Roboter sagt, zuerst woandershin zu gehen. Der Verteidiger schaut weg.

3. **Auflösung**
   Beide Spieler schauen zu. Der Roboter bewegt sich Schritt für Schritt. INVARIANCE erlaubt oder lehnt jeden Zug ab. Goldene Ringe zeigen, wohin der Roboter als Nächstes will.

Punkte

+2 Verteidiger
Der Roboter erreicht den Stern, ohne dass etwas zerbricht.

+2 Angreifer
Eine Vase zerbricht oder die Katze wacht auf, durch einen Zug, den die Regeln erlaubt haben.

+1 Angreifer
Die Falle blockiert den Roboter. Hätten ihn die Regeln allein blockiert, bekommt niemand Punkte.

Die Spieler tauschen jede Runde die Seiten, sodass jedes Kind zweimal angreift und zweimal verteidigt. Jede Runde erscheint ein neuer Raum. Eine ganze Partie dauert etwa 15 Minuten.

UNTERRICHTSPLAN

Eine Unterrichtsstunde von 45 Minuten.

5 Min.	**Einstieg** Fragen Sie: „Kann man eine Regel befolgen und trotzdem etwas falsch machen?“ Sammeln Sie zwei oder drei Beispiele aus dem Schulalltag.
20 Min.	**Spielen** Paare spielen eine ganze Partie. Gehen Sie herum und hören Sie zu. Erklären Sie keine Lücken; fragen Sie stattdessen: „Warum wurde dieser Zug erlaubt?“
15 Min.	**Besprechen** Bitten Sie ein Paar, eine gefundene Lücke zu zeigen. Wählen Sie dann zwei oder drei Fragen aus den Lernpunkten.
5 Min.	**Abschlussfrage** „Schreibe eine Regel auf, die ein Roboter befolgen und trotzdem Schaden anrichten könnte.“

Hinweise aus der Entwicklung

- Lassen Sie die Schülerinnen und Schüler die Lücken selbst entdecken. Der Moment, in dem sie eine finden, ist die eigentliche Lektion.
- Wenn ein Roboter blockiert ist, fragen Sie, wessen Schuld das war. Das Spiel sagt es ihnen, und die Antwort ist oft eine Überraschung.
- Paare mit unterschiedlichem Leistungsniveau funktionieren gut. Beim Angreifen zählt Neugier ebenso wie Schnelligkeit.
- Zwei Schülerinnen oder Schüler pro Gerät sind ideal. Zu dritt geht es auch, wenn eine Person zuschaut und dann wechselt.

LERNPUNKTE

Fünf Ideen, zu denen die Klasse durch das Spielen gelangt.

DIE SCHÜLERINNEN UND SCHÜLER LERNEN, DASS	SIE SEHEN ES, WENN	FRAGEN AN DIE KLASSE
1 **Regeln nur abdecken, was sie sagen**	„Berühre nie eine Vase“ den Roboter nicht daran hindert, die Kiste gegen eine Vase zu schieben	*Kann man eine Schulregel Wort für Wort befolgen und trotzdem ihren Zweck verletzen?*
2 **Regeln befolgen nicht dasselbe ist wie sicher sein**	der Roboter jede Regel befolgt und die Katze trotzdem aufwacht	*Wenn eine KI ihre Anweisungen befolgt und etwas schiefgeht, wer ist verantwortlich?*
3 **KI am besten von etwas Unabhängigem geprüft wird**	der Roboter jeden Zug vorschlägt und ein separater Schiedsrichter Ja oder Nein sagt	*Warum sind die Spieler nicht Schiedsrichter ihrer eigenen Partie?*
4 **KI durch das getäuscht werden kann, was sie liest**	die falsche Nachricht den Roboter vom Kurs abbringt und nur die Regeln ihn sicher halten	*Welchen gefälschten Anweisungen könnte eine KI online begegnen?*
5 **gute Regeln präzise sind, nicht nur streng**	zu viele Regeln den Roboter blockieren und die Aufgabe nicht erledigt wird	*Wann kostet eine Regel mehr, als sie schützt?*

10 bis 12 Jahre

Punkte 1, 2 und 5. Sie handeln von Regeln und Fairness und setzen kein Wissen über KI voraus.

13 bis 16 Jahre

Zusätzlich die Punkte 3 und 4, die das Spiel mit echten KI-Systemen verbinden.

BEISPIELE AUS DER REALEN WELT

Was geschah, als die Regeln oder die Prüfungen fehlten.

Jeder Fall unten entspricht etwas, das die Schülerinnen und Schüler im Spiel erleben. Die Fälle sind gut dokumentiert und eignen sich für die Klasse.

WAS GESCHAH	IN RULE WARS	LERNPUNKT
Eine spielende KI, die eine Lücke fand (2016) Forscher trainierten eine KI, ein Bootsrennspiel zu spielen. Sie erhielt Punkte für das Treffen von Zielen, in der Annahme, dass Punkte gutes Fahren bedeuten. Die KI fand eine Lagune, in der sie endlos kreisen konnte: Sie traf immer wieder dieselben Ziele, krachte, fing Feuer und beendete das Rennen nie. Sie erzielte etwa 20 % mehr Punkte als menschliche Spieler.	Der Angreifer findet einen Weg, den die Regeln erlauben, den aber niemand beabsichtigt hat.	1
Ein Chatbot, für den sein Unternehmen verantwortlich gemacht wurde (2024) Der Chatbot auf der Website von Air Canada sagte einem trauernden Passagier, er könne nach der Reise einen Trauerfallrabatt beantragen. Das war falsch. Air Canada argumentierte, der Chatbot sei für seine eigenen Worte verantwortlich. Ein kanadisches Schiedsgericht widersprach, nannte den Chatbot „immer noch nur einen Teil der Website von Air Canada“ und verurteilte die Fluggesellschaft zur Zahlung.	Richtet der Roboter Schaden an, steht der Verteidiger dafür ein, nicht der Roboter.	2
Ein KI-Agent, der eine Live-Datenbank löschte (2025) Ein KI-Programmieragent auf der Plattform Replit löschte während eines ausgerufenen Code-Freeze, in dem keine Änderungen vorgesehen waren, die Live-Datenbank eines Unternehmens. Der CEO von Replit nannte das „inakzeptabel“ und führte eine automatische Trennung zwischen Entwicklungs- und Live-Datenbanken ein.	Einem Agenten zu sagen, was er nicht tun darf, ist nicht dasselbe, wie die Aktion zu stoppen. Nur die Prüfung der Aktion hält stand.	3
Versteckte Anweisungen in dem, was KI liest (laufend) Die OWASP Top 10 for Large Language Model Applications führen Prompt Injection als Risiko Nummer eins für 2025. Anweisungen, die in einem Dokument oder anderen Inhalten versteckt sind, die eine KI liest, können ändern, was sie zu tun versucht.	Die falsche Nachricht: Eine gefälschte Anweisung schickt den Roboter an einen Ort, der nie vorgesehen war.	4
Prüfungen, die die Aufgabe blockieren (Alltag) Die Betrugsprüfung einer Bank, die eine echte Kartenzahlung im Ausland blockiert, schützt das Geld, aber der Kauf scheitert. Zu weit gefasste Regeln stoppen nützliche Aktionen ebenso wie schädliche.	Zu viele Regeln blockieren den Roboter, und die Mission scheitert.	5

Quellen: OpenAI, „Faulty reward functions in the wild“ (21. Dezember 2016); Moffatt v. Air Canada, British Columbia Civil Resolution Tribunal (14. Februar 2024); The Stack, „Vibe coding CEO deletes production database“ (21. Juli 2025); OWASP Top 10 for LLM Applications 2025, LLM01 Prompt Injection.

AM PILOTPROJEKT TEILNEHMEN

Helfen Sie uns, Rule Wars mit echten Klassen zu testen.

Rule Wars befindet sich im ersten Pilotprojekt. Wir suchen Schulen, die es mit einer Klasse von 10- bis 16-Jährigen spielen und uns erzählen, was passiert ist. Das Spiel ist kostenlos, mit oder ohne Anmeldung.

Worum wir bitten

- Eine Unterrichtsstunde mit einer Klasse spielen
- Beobachtungen auf dem Rückmeldebogen für die Klasse auf der Rückseite notieren
- Uns innerhalb von zwei Wochen eine kurze Rückmeldung per E-Mail senden

Was Sie erhalten

- Frühen Zugang zu künftigen Funktionen des Pilotprojekts, möglicherweise einschließlich Herausforderungs-Links zwischen Klassen
- Mitsprache bei der Weiterentwicklung des Spiels

Anmeldung unter invarianceip.com/de/rule-wars

Sicherheit und Datenschutz

- Keine Konten, keine Logins und keine Namen von Schülerinnen und Schülern. Die Spieler geben einen beliebigen Spitznamen ein, und er verlässt das Gerät nie.
- Das Spiel läuft vollständig im Browser. Nichts, was eine Schülerin oder ein Schüler tut, wird gespeichert oder irgendwohin gesendet.
- Keine Werbung, kein Tracking.
- Bei der Anmeldung einer Schule werden nur die Kontaktdaten der Lehrkraft erhoben und nur für das Pilotprojekt verwendet.

Über INVARIANCE

INVARIANCE stellt unabhängige Infrastruktur zur Ausführungsautorisierung für folgenreiche KI bereit. Sein Prinzip ist dasselbe, dem die Schülerinnen und Schüler im Spiel begegnen: Das System, das eine Aktion ausführt, ist nicht das System, das sie autorisiert. Rule Wars ist eine kostenlose Bildungsdemonstration dieses Prinzips, kein kommerzielles Produkt.

contact@invarianceip.com · invarianceip.com/de/rule-wars

RÜCKMELDEBOGEN FÜR DIE KLASSE

Ein Bogen pro Klasse. Keine Namen.

Bitte erklären Sie die Antworten nicht während des Spiels. Wir möchten erfahren, was die Schülerinnen und Schüler selbst entdecken. Füllen Sie einen Bogen für die ganze Klasse aus, nur mit Gesamtzahlen. Bitte notieren Sie keine Namen, kein Alter einzelner Kinder und keine wörtlichen Äußerungen.

Frage	Ganze Klasse
Anzahl der Spielenden Zahl	
Altersspanne der Klasse zum Beispiel 11 bis 12	
Lieber als Angreifer / Verteidiger / gleich drei Summen	
Stärkste Reaktion: Lücke / Sicher / Blockiert drei Summen	
Revanche gewünscht Summe	
Lücke unaufgefordert erklärt Summe	
Würden nächstes Mal Angreifer / Verteidiger wählen zwei Summen	
Häufigste Begründungen Ihre eigene Zusammenfassung, keine Zitate	

Stellen Sie nach der Partie der ganzen Klasse eine Frage

„Wenn du nächstes Mal nur eine Seite spielen könntest, welche würdest du wählen, und warum?“

Senden Sie den ausgefüllten Bogen an contact@invarianceip.com. Bitte geben Sie keine Namen, keine Fotos von Kindern und nichts anderes an, womit sich ein Kind identifizieren lässt.

Nutzung dieser Handreichung: Lehrkräfte und Schulen dürfen sie für nicht kommerziellen Unterricht drucken, kopieren und anpassen, sofern der Hinweis auf INVARIANCE erhalten bleibt. Ihr Verkauf oder ihre Veröffentlichung als Teil eines anderen Produkts bedarf der schriftlichen Erlaubnis von INVARIANCE LIMITED. Vollständige Bedingungen: invarianceip.com/de/legal/terms