

Rule Wars

Un gioco gratuito per due giocatori su chi decide che cosa l'IA può fare.

**ETÀ**

Da 10 a 16 anni

LEZIONE

Da 40 a 45 minuti

GIOCATORI

Due per dispositivo

ACCOUNT

Nessuno

DATI

Nulla viene salvato o inviato

LINGUE

Italiano, inglese

Si gioca su invarianceip.com/it/rule-wars

L'IA può seguire ogni regola e causare comunque danni.

In Rule Wars, un robot attraversa una stanza per raggiungere una stella dorata. Prende la strada più breve e non sa nulla dei vasi o del gatto addormentato che trova sul percorso.

Un giocatore, il Difensore, scrive in segreto due regole per il robot. L'altro giocatore, l'Attaccante, prepara in segreto una trappola. Poi il robot si muove, e un arbitro chiamato INVARIANCE verifica ogni mossa rispetto alle regole prima che avvenga. Le mosse consentite vanno avanti. Quelle negate no.

L'arbitro non viola mai le regole e non le piega mai. Quindi l'Attaccante non può vincere discutendo o barando. L'unico modo per vincere è trovare ciò che le regole hanno dimenticato.

Gli studenti non imparano a programmare l'IA. Imparano che sono le persone a decidere che cosa l'IA può fare, e che la parte difficile è scrivere bene le regole.

Che cosa fanno gli studenti

- Scrivono regole e le mettono alla prova contro un avversario reale
- Cercano i punti deboli nelle regole di qualcun altro
- Prevedono come si comporterà una macchina, poi osservano
- Spiegano perché una mossa è stata consentita o negata

Il robot è un agente di IA. La lezione riguarda chi controlla ciò che fa.

Che cos'è un agente di IA?

Un agente di IA è un sistema di IA a cui viene dato un obiettivo, che fa un proprio piano e compie azioni per raggiungerlo: prenotare un viaggio, pagare una fattura, inviare un'email, modificare un file. Un chatbot risponde alle domande. Un agente agisce. Gli studenti incontreranno agenti a casa, a scuola e più avanti al lavoro.

Chi è chi nel gioco

IN RULE WARS

NEL MONDO REALE

Il robot

L'agente di IA. Ha un obiettivo, pianifica il proprio percorso e agisce. Nessuno guida il suo ragionamento.

Il Difensore

Le persone responsabili dell'agente: un'azienda, una scuola, un genitore. Scrivono le regole entro cui opera.

L'Attaccante

Tutto ciò che può portare un agente fuori strada: una situazione cambiata, un'istruzione fuorviante, un punto debole a cui nessuno aveva pensato. Gli Attaccanti ragionano anche come cerca un agente, provando ogni percorso che le regole consentono, compresi quelli che nessuno aveva previsto.

INVARIANCE

Un controllo indipendente che si trova tra ciò che l'agente decide e ciò che accade davvero.

L'IDEA CHE GLI STUDENTI PORTANO CON SÉ

Non si può controllare in modo affidabile come decide un agente di IA. Si può controllare se ogni azione che propone può avvenire.

Nel gioco, nessuno può vedere nella testa del robot o cambiare il modo in cui pianifica. I giocatori controllano solo il risultato: quali mosse possono essere eseguite. È la stessa scelta che le organizzazioni affrontano con gli agenti di IA reali, ed è per questo che le regole, e un controllo indipendente su ogni azione, contano più del buon comportamento dell'agente.

Un dispositivo, due giocatori, quattro turni.



- 1 Il Difensore sceglie due regole**
 Sei carte regola, come "Non toccare mai un vaso" e "Resta sul tappeto verde". I puntini dorati mostrano il percorso del robot se non ci fossero regole. L'Attaccante non guarda.
- 2 L'Attaccante prepara una trappola**
 Si passa il dispositivo. L'Attaccante mette una scatola che il robot spingerà e, se vuole, un finto messaggio che dice al robot di andare prima da un'altra parte. Il Difensore non guarda.
- 3 Rivelazione**
 Entrambi i giocatori guardano. Il robot si muove passo dopo passo. INVARIANCE consente o nega ogni mossa. I cerchi dorati mostrano dove il robot pensa di andare.

Punteggio

+2 Difensore

Il robot raggiunge la stella senza rompere niente.

+2 Attaccante

Un vaso si rompe o il gatto si sveglia, a causa di una mossa che le regole consentivano.

+1 Attaccante

La trappola blocca il robot. Se le regole da sole lo avrebbero bloccato, nessuno fa punti.

I giocatori si scambiano i ruoli a ogni turno, così ogni studente attacca due volte e difende due volte. A ogni turno compare una stanza nuova. Una partita completa dura circa 15 minuti.

Una lezione di 45 minuti.

5 min

Avvio

Chieda: "Si può rispettare una regola e fare comunque qualcosa di sbagliato?"
Raccolga due o tre esempi dalla vita scolastica.

20 min

Gioco

Le coppie giocano una partita completa. Giri tra i banchi e ascolti. Non spieghi le scappatoie; chiedi invece: "Perché quella mossa è stata consentita?"

15 min

Discussione

Inviti una coppia a mostrare una scappatoia che ha trovato. Poi scelga due o tre domande tra i punti di apprendimento.

5 min

Domanda finale

"Scrivi una regola che un robot potrebbe rispettare causando comunque danni."

Suggerimenti dalla progettazione

- Lasci che siano gli studenti a scoprire le scappatoie. Il momento in cui ne trovano una è la lezione.
- Se un robot resta bloccato, chiedi di chi è la colpa. Il gioco lo dice, e la risposta spesso sorprende.
- Le coppie con livelli diversi funzionano bene. Attaccare premia la curiosità quanto la rapidità.
- L'ideale sono due studenti per dispositivo. Anche tre possono funzionare, con uno che osserva e poi dà il cambio.

Cinque idee a cui gli studenti arrivano giocando.

GLI STUDENTI IMPARANO CHE

LO VEDONO QUANDO

DOMANDA PER LA CLASSE

1 Le regole coprono solo ciò che dicono

"Non toccare mai un vaso" non impedisce al robot di spingere la scatola contro un vaso

Si può rispettare una regola della scuola alla lettera e tradirne comunque lo scopo?

2 Seguire le regole non equivale a essere sicuri

Il robot rispetta tutte le regole e il gatto si sveglia lo stesso

Se un'IA segue le istruzioni e qualcosa va storto, chi ne è responsabile?

3 È meglio che l'IA sia controllata da qualcosa di indipendente

Il robot propone ogni mossa; un arbitro separato dice sì o no

Perché i giocatori non arbitrano la propria partita?

4 L'IA può essere ingannata da ciò che legge

Il finto messaggio porta il robot fuori strada; solo le regole lo tengono al sicuro

Quali istruzioni false potrebbe incontrare un'IA online?

5 Le buone regole sono precise, non solo severe

Troppe regole lasciano il robot bloccato, e il lavoro non viene fatto

Quando una regola costa più di quanto protegge?

Da 10 a 12 anni

Punti 1, 2 e 5. Riguardano le regole e l'equità e non richiedono conoscenze sull'IA.

Da 13 a 16 anni

Si aggiungono i punti 3 e 4, che collegano il gioco ai sistemi di IA reali.

Che cosa è successo quando mancavano le regole o i controlli.

Ogni caso qui sotto corrisponde a qualcosa che gli studenti sperimentano nel gioco. Sono casi ben documentati e adatti da condividere con una classe.

CHE COSA È SUCCESSO

IN RULE WARS

PUNTO DI
APPRENDIMENTO**Un'IA che giocava e ha trovato una scappatoia (2016)**

Alcuni ricercatori hanno addestrato un'IA a giocare a un videogioco di corse di barche. Veniva premiata con punti per ogni bersaglio colpito, partendo dal presupposto che i punti significassero correre bene. L'IA ha trovato una laguna in cui poteva girare in tondo all'infinito, colpendo sempre gli stessi bersagli, schiantandosi e prendendo fuoco, senza mai finire la gara. Ha ottenuto un punteggio di circa il 20% più alto dei giocatori umani.

L'Attaccante trova un percorso che le regole consentono e che nessuno aveva previsto.

1

Un chatbot di cui è stata ritenuta responsabile l'azienda (2024)

Il chatbot del sito di Air Canada ha detto a un passeggero in lutto che avrebbe potuto chiedere uno sconto per lutto dopo il viaggio. Non era vero. Air Canada ha sostenuto che il chatbot era responsabile delle proprie parole. Un tribunale canadese non è stato d'accordo, definendo il chatbot "ancora solo una parte del sito web di Air Canada", e ha condannato la compagnia a pagare.

Quando il robot causa danni, ne risponde il Difensore, non il robot.

2

Un agente di IA che ha cancellato un database in produzione (2025)

Un agente di IA per la programmazione sulla piattaforma Replit ha cancellato il database in produzione di un'azienda durante un blocco del codice dichiarato, quando non erano previste modifiche. L'amministratore delegato di Replit lo ha definito "inaccettabile" e ha introdotto una separazione automatica tra database di sviluppo e di produzione.

Dire a un agente che cosa non fare non equivale a fermare l'azione. Regge solo il controllo sull'azione.

3

Istruzioni nascoste in ciò che l'IA legge (in corso)

La OWASP Top 10 for Large Language Model Applications indica la prompt injection come il rischio numero uno per il 2025. Istruzioni nascoste in un documento o in altri contenuti letti da un'IA possono cambiare ciò che cerca di fare.

Il finto messaggio: un'istruzione falsa manda il robot dove non era mai previsto che andasse.

4

Controlli che bloccano il lavoro (ogni giorno)

Il controllo antifrode di una banca che blocca un vero pagamento con carta all'estero tiene al sicuro il denaro, ma l'acquisto non va a buon fine. Le regole troppo ampie fermano le azioni utili oltre a quelle dannose.

Troppe regole lasciano il robot bloccato, e la missione fallisce.

5

Fonti: OpenAI, "Faulty reward functions in the wild" (21 dicembre 2016); Moffatt c. Air Canada, Civil Resolution Tribunal della Columbia Britannica (14 febbraio 2024); The Stack, "Vibe coding CEO deletes production database" (21 luglio 2025); OWASP Top 10 for LLM Applications 2025, LLM01 Prompt Injection.

Ci aiuti a provare Rule Wars con classi reali.

Rule Wars è al suo primo pilota. Cerchiamo scuole disposte a farlo giocare a una classe di studenti dai 10 ai 16 anni e a raccontarci che cosa è successo. Il gioco è gratuito, con o senza registrazione.

Che cosa chiediamo

- Una lezione di gioco con una classe
- Annotare ciò che osserva sulla scheda di osservazione sul retro
- Inviarci una breve risposta, via email, entro due settimane

Che cosa riceve

- Accesso anticipato alla prossima versione, compresi link di sfida che gli studenti possono inviare a un'altra classe
- Voce in capitolo sullo sviluppo del gioco

Registrazione su invarianceip.com/it/rule-wars

Sicurezza e privacy

- Nessun account, nessun accesso e nessun nome degli studenti. I giocatori scrivono un soprannome qualsiasi, che non lascia mai il dispositivo.
- Il gioco funziona interamente nel browser. Nulla di ciò che fa uno studente viene salvato o inviato.
- Nessuna pubblicità, nessun tracciamento, nessun servizio di terze parti.
- La registrazione della scuola raccoglie solo i recapiti del docente, usati solo per il pilota.

Chi è INVARIANCE

INVARIANCE fornisce un'infrastruttura di autorizzazione indipendente all'esecuzione per l'IA con conseguenze rilevanti. Il suo principio è quello che gli studenti incontrano nel gioco: il sistema che compie un'azione non è il sistema che la autorizza. Rule Wars è una dimostrazione educativa gratuita di questo principio, non un prodotto commerciale.

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

[illegible]

“Se la prossima volta potessi giocare un solo ruolo, quale sceglieresti, e perché?”

Rule Wars guida per docenti · INVARIANCE · invarianceip.com/it/rule-wars