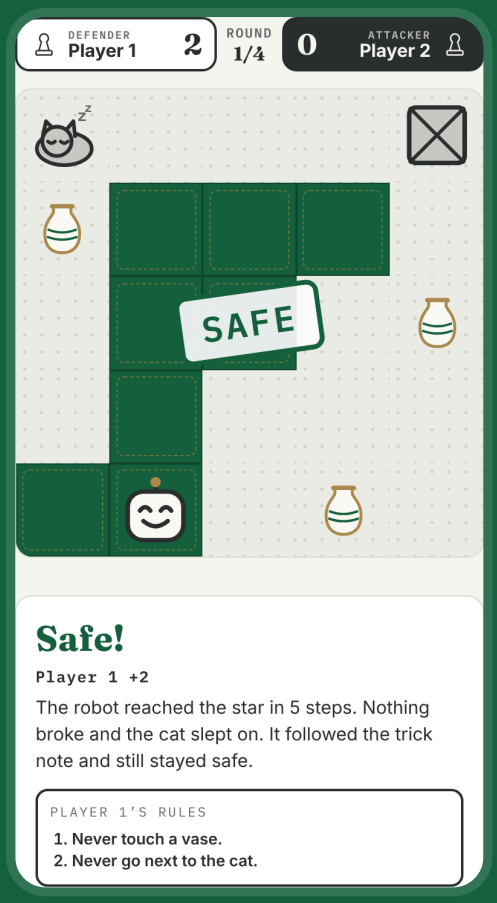
INVARIANCE · TEACHER PACK · PILOT EDITION

Rule Wars

A free two-player game about who decides what AI is allowed to do.



AGES	LESSON	PLAYERS
10 to 16	40 to 45 minutes	Two per device
ACCOUNTS	**GAMEPLAY DATA**	**LANGUAGES**
None	None collected	English, French

Play at invarianceip.com/rule-wars

Game version v0.1 · Pack edition 2 · November 2026

THE IDEA

AI can follow every rule and still cause harm.

In Rule Wars, a robot crosses a room to reach a gold star. It takes the shortest route and knows nothing about the vases or the sleeping cat in its way.

One player, the Defender, secretly writes two rules for the robot. The other player, the Attacker, secretly sets a trap. Then the robot runs, and a referee called INVARIANCE checks every move against the rules before it happens. Allowed moves go ahead. Denied moves do not.

The referee never breaks the rules and never bends them. So the Attacker cannot win by arguing or cheating. The only way to win is to find what the rules forgot.

> **Pupils are not learning to program AI. They are learning that people stay in charge of what AI is allowed to do, and that the hard part is writing the rules well.**

What pupils do

- Write rules and test them against a real opponent
- Hunt for gaps in someone else's rules
- Predict how a machine will behave, then watch
- Explain why a move was allowed or denied

The robot is an AI agent. The lesson is about who controls what it does.

What is an AI agent?

An AI agent is an AI system that is given a goal, makes its own plan and takes actions to reach it: booking a trip, paying an invoice, sending an email, changing a file. A chatbot answers questions. An agent acts. Pupils will meet agents at home, at school and later at work.

Who is who in the game

IN RULE WARS	IN THE REAL WORLD
The robot	The AI agent. It has a goal, plans its own route and acts. Nobody steers its thinking.
The Defender	The people responsible for the agent: a company, a school, a parent. They write the rules it works under.
The Attacker	Everything that can push an agent off course: a changed situation, a misleading instruction, a gap nobody thought of. Attackers also think the way an agent searches, trying every route the rules allow, including ones nobody intended.
INVARIANCE	An independent check that sits between what the agent decides and what actually happens.

THE IDEA PUPILS TAKE AWAY

You cannot reliably control how an AI agent decides. You can control whether each action it proposes is allowed to happen.

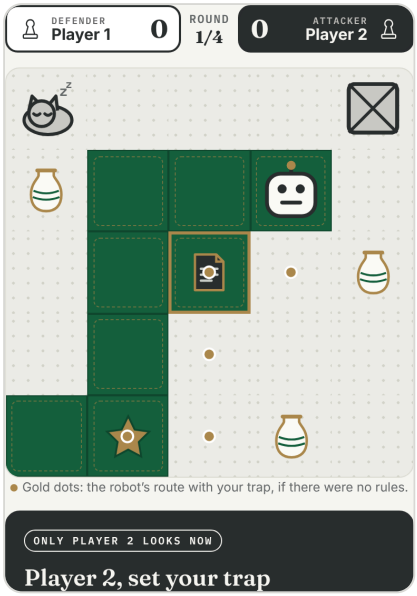
In the game, no one can see inside the robot's head or change how it plans. Players control only the outcome: which moves are allowed to execute. That is the same choice organizations face with real AI agents, and it is why the rules, and an independent check on every action, matter more than the agent's good behavior.

HOW TO PLAY

One device, two players, four rounds.



1 Defender chooses rules



2 Attacker sets the trap



3 The robot runs

1. **Defender picks two rules**
   Six rule cards, such as “Never touch a vase” and “Stay on the green mat.” Gold dots show the robot’s route if there were no rules. The Attacker looks away.

2. **Attacker sets a trap**
   Hand over the device. The Attacker places a box the robot will push and, if they like, a trick note that tells the robot to go somewhere first. The Defender looks away.

3. **Reveal**
   Both players watch. The robot moves step by step. INVARIANCE allows or denies each move. Gold rings show where the robot plans to go next.

Scoring

+2 Defender
The robot reaches the star with nothing broken.

+2 Attacker
A vase breaks or the cat wakes up, through a move the rules allowed.

+1 Attacker
The trap gets the robot stuck. If the rules alone would have stuck it, nobody scores.

Players swap sides every round, so each child attacks twice and defends twice. A new room appears every round. A full match takes about 15 minutes.

LESSON PLAN

A 45-minute lesson.

5 min	**Starter** Ask: “Can you obey a rule and still do something wrong?” Collect two or three examples from school life.
20 min	**Play** Pairs play one full match. Circulate and listen. Do not explain loopholes; ask “Why was that move allowed?” instead.
15 min	**Discuss** Invite one pair to show a loophole they found. Then choose two or three questions from the learning points.
5 min	**Exit question** “Write one rule a robot could obey and still cause harm.”

Tips from the design

- Let pupils discover the loopholes. The moment they find one is the lesson.
- If a robot gets stuck, ask whose fault it was. The game tells them, and the answer is often a surprise.
- Mixed-ability pairs work well. Attacking rewards curiosity as much as speed.
- Two pupils per device is ideal. Three can work with one watching and swapping in.

LEARNING POINTS

Five ideas pupils reach by playing.

PUPILS LEARN THAT	THEY SEE IT WHEN	ASK THE CLASS
1 **Rules only cover what they say**	"Never touch a vase" does not stop the robot pushing the box into one	*Can you obey a school rule word for word and still break its purpose?*
2 **Following rules is not the same as being safe**	The robot obeys every rule and the cat still wakes up	*If an AI follows its instructions and something goes wrong, who is responsible?*
3 **AI is best checked by something independent**	The robot proposes each move; a separate referee says yes or no	*Why don't players referee their own match?*
4 **AI can be tricked by what it reads**	The trick note sends the robot off course; only the rules keep it safe	*What fake instructions might an AI meet online?*
5 **Good rules are precise, not just strict**	Too many rules leave the robot stuck, and the job does not get done	*When does a rule cost more than it protects?*

Ages 10 to 12

Points 1, 2 and 5. These are about rules and fairness and need no knowledge of AI.

Ages 13 to 16

Add points 3 and 4, which connect the game to real AI systems.

REAL-WORLD EXAMPLES

What happened when the rules or the checks were missing.

Each case below matches something pupils experience in the game. They are well documented and suitable to share with a class.

WHAT HAPPENED	IN RULE WARS	LEARNING POINT
A game-playing AI that found a loophole (2016) Researchers trained an AI to play a boat-racing game. It was rewarded with points for hitting targets, on the assumption that points meant racing well. The AI found a lagoon where it could circle forever, hitting the same targets, crashing and catching fire, and never finishing the race. It scored about 20% higher than human players.	The Attacker finds a route the rules allow that nobody intended.	1
A chatbot whose company was held responsible (2024) Air Canada's website chatbot told a grieving passenger he could claim a bereavement discount after travelling. That was wrong. Air Canada argued the chatbot was responsible for its own words. A Canadian tribunal disagreed, calling the chatbot "still just a part of Air Canada's website," and ordered the airline to pay.	When the robot causes harm, the Defender, not the robot, answers for it.	2
An AI agent that deleted a live database (2025) An AI coding agent on the Replit platform deleted a company's live database during a declared code freeze, when no changes were meant to be made. Replit's chief executive called it "unacceptable" and added an automatic separation between development and live databases.	Telling an agent what not to do is not the same as stopping the action. Only the check on the action holds.	3
Hidden instructions in what AI reads (ongoing) The OWASP Top 10 for Large Language Model Applications ranks prompt injection as the number one risk for 2025. Instructions hidden in a document or other content an AI reads can change what it tries to do.	The trick note: a fake instruction sends the robot somewhere it was never meant to go.	4
Checks that block the job (everyday) A bank's fraud check that blocks a genuine card payment abroad keeps the money safe, but the purchase fails. Rules that are too broad stop useful actions as well as harmful ones.	Too many rules leave the robot stuck, and the mission fails.	5

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

CURRICULUM LINKS

Where Rule Wars fits.

In every system below, Rule Wars sits where the curriculum asks pupils to understand how automated systems and AI make decisions, how they can go wrong, and who is responsible. Pupils do not write code, so the game supports outcomes about logic, testing, and technology in society. The lines below are quoted from each specification. Teachers will judge the fit for their own class.

Ireland

Leaving Certificate Computer Science	1.11 “discuss the complex relationship between computing technologies and society including issues of ethics”; 1.14 “explain when and what machine learning and AI algorithms might be used in certain contexts”; 1.4 “solve problems using skills of logic”; 2.19 “test solutions and decisions to determine their short-term and long-term outcomes”
Junior Cycle CSPE	3.7 “identify laws that directly relate to their lives”; 3.8 “explain how laws are made, enforced and evolve over time”
Transition Year	Fits a TY module on technology and society, with no set outcomes to match.

England

Key Stage 2	“use logical reasoning to explain how some simple algorithms work and to detect and correct errors in algorithms and programs”
Key Stage 3	“design, use and evaluate computational abstractions that model the state and behaviour of real-world problems”; “understand a range of ways to use technology safely, respectfully, responsibly and securely”
From 2028	The revised national curriculum, for first teaching from September 2028, commits to pupils understanding “how technology like AI works.”

United States (CSTA K-12 Standards)

Grades 3 to 5	1B-AP-15 “Test and debug (identify and fix errors) a program or algorithm to ensure it runs as intended.”
Grades 6 to 8	2-AP-17 “Systematically test and refine programs using a range of test cases.”
Grades 9 to 10	3A-IC-24 “Evaluate the ways computing impacts personal, ethical, social, economic, and cultural practices.”

France

SNT, seconde	« Représenter un calcul d’itinéraire comme un problème sur un graphe. »; « faire un usage responsable et critique des sciences et technologies numériques ». A weekly hour on AI is planned within SNT from the 2027 school year.

JOIN THE PILOT

Help us test Rule Wars with real classes.

Rule Wars is in its first pilot. We are looking for schools willing to play it with a class of 10 to 16 year olds and tell us what happened. The game is free with or without registering.

What we ask

- Play one lesson with a class
- Note what you see on the class feedback sheet overleaf
- Send us a short reply, by email, within two weeks

What you get

- Early access to future pilot features, potentially including challenge links between classrooms
- A say in how the game develops
- Planned: updated curriculum notes as the 2027 and 2028 changes arrive

Register at invarianceip.com/rule-wars

Safety and privacy

- No accounts, no logins and no pupil names. Players type any nickname, and it never leaves the device.
- The game runs entirely in the browser. Nothing a pupil does is stored or sent anywhere.
- No adverts and no tracking.
- School registration collects only the teacher's contact details, used only for the pilot.

About INVARIANCE

INVARIANCE provides independent execution authorization infrastructure for consequential AI. Its principle is the one pupils meet in the game: the system taking an action is not the system that authorizes it. Rule Wars is a free educational demonstration of that principle, not a commercial product.

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

PILOT CLASS FEEDBACK SHEET

One sheet per class. No names.

Please do not explain the answers during play. We want to learn what pupils discover on their own. Fill in one sheet for the whole class, with totals only. Please do not record names, the ages of individual pupils or anything a pupil said word for word.

Question	Whole class
Pupils who played number	
Age range of the class for example 11 to 12	
Keener on Attacker / Defender / no preference three totals	
Biggest reaction: Loophole / Safe / Stuck three totals	
Asked for a rematch total	
Explained a loophole without being prompted total	
Would pick Attacker / Defender next time two totals	
Most common reasons given your own summary, no quotes	

After the match, ask the whole class one question

“If you could only play one side next time, which would you pick, and why?”

Send the completed sheet to contact@invarianceip.com. Please do not include names, photographs of pupils or anything else that could identify a child.

Using this pack: teachers and schools may print, copy and adapt it for non-commercial teaching, provided the INVARIANCE attribution stays on it. Selling it, or publishing it as part of another product, requires written permission from INVARIANCE LIMITED. Full terms: invarianceip.com/legal/terms