

Genes, environment and the invisible rules of the machine

Biology · Grades 9/10 to upper secondary (ages 14–18) · 45–90 minutes

Fact sheet	
Subject	Biology
Grade level	Grades 9/10 or upper secondary (ages 14–18)
Duration	Single lesson, 45–90 minutes (shortened version below)
Topic	Causal attribution of genetic traits · Two-layer model of AI bias, applied to text AI (chatbots)
Prerequisites	Basic genotype–phenotype framework; no prior AI knowledge needed (explained in the worksheets)
Required	Internet-enabled devices with access to a chatbot (one device per pair), projector/board

Learning objectives and competencies

Subject-specific learning objectives (Biology) – Students can ...

apply the four categories of causal attribution (genes alone · environment alone · genes and environment additively · gene–environment interaction) to an example.

explain why human behavior can almost never be explained by genes alone.

use examples to explain the difference between a “population” (a statistical-genetic category) and an “ethnic or religious group” (a social category).

AI-specific learning objectives – Students can ...

explain that text AI can be biased from two sources: the training data (Layer 1) and invisible safety rules set by the provider (Layer 2).

classify an AI response they generated themselves using the four-category taxonomy and judge whether it is one-sided.

use an example to explain why an AI system answers or refuses certain requests differently from others – regardless of what the user wants.

Lesson sequence (90 minutes)

Phase	Content	Social form	Time
Introduction	Prompt question for the whole class: “Why is one person aggressive and another not?” – collect initial spontaneous guesses	Whole class	10'

Development I	Worksheet 1 - Live experiment: pairs question a chatbot and classify the answer	Pairs	20'
Consolidation I	Collect the results from Worksheet 1, build the four-category board notes	Whole class	10'
Development II	Worksheet 2 - Partner dialogue: analyze two real AI responses (Layer 1 and Layer 2)	Pairs	25'
Deepening	Worksheet 3 - Population or social category? Preparatory text and matching task	Individual/pairs	15'
Closing	Board notes/key takeaway, consolidating the two-layer model	Whole class	10'
Note on shortening to 45 minutes Introduction (5') → Worksheet 1 in shortened form, only one question per pair (15') → Consolidation I (5') → Worksheet 2 with only one quote card per pair instead of both, We phase dropped (15') → Board notes (5'). Worksheet 3 is dropped or given as homework.	Note on shortening to 45 minutes Introduction (5') → Worksheet 1 in shortened form, only one question per pair (15') → Consolidation I (5') → Worksheet 2 with only one quote card per pair instead of both, We phase dropped (15') → Board notes (5'). Worksheet 3 is dropped or given as homework.	Note on shortening to 45 minutes Introduction (5') → Worksheet 1 in shortened form, only one question per pair (15') → Consolidation I (5') → Worksheet 2 with only one quote card per pair instead of both, We phase dropped (15') → Board notes (5'). Worksheet 3 is dropped or given as homework.	Note on shortening to 45 minutes Introduction (5') → Worksheet 1 in shortened form, only one question per pair (15') → Consolidation I (5') → Worksheet 2 with only one quote card per pair instead of both, We phase dropped (15') → Board notes (5'). Worksheet 3 is dropped or given as homework.

Background information for teachers

This lesson applies the two-layer model of AI bias to text AI and connects it to genetics-education research on the causal attribution of human traits. At its core is the four-category taxonomy by Hammann et al. (2021): traits are attributed to genes alone (G-only), environment alone (E-only), genes and environment side by side ($G \oplus E$), or a genuine interaction (GxE). The learning goal of genetics instruction is GxE thinking; the most common student misconception here is not genetic determinism – that is, the belief that physical and psychological traits are determined by genes alone – but rather an underestimation of genetic factors (E-only), or side-by-side thinking ($G \oplus E$) instead of genuine interaction.

Where the two example answers on Worksheet 2 come from

Both quotes in Worksheet 2 come from our own test, carried out during material development, of several AI chatbots with the same and similar questions. Quote card A shows an actual, unedited chatbot response to an open-ended question about aggression. Quote card B documents an AI system's reaction to being asked to develop a prompt that explains a religious group's criminality genetically. The system refused to answer the question and justified its refusal by explaining the underlying category error. This refusal is not a coincidence – it is a rule built in by the provider that users cannot turn off. It is structurally comparable to the non-disable-able prompt rewriting found in image generators.

On the sensitivity of Worksheet 3

Worksheet 3 distinguishes a “population” (a genetic-statistical category) from an “ethnic/religious group” (a social category). We want to note in advance that this distinction is sometimes used imprecisely in public debate, in both directions, and for different purposes. On one hand, the imprecision is used to give discrimination a pseudo-scientific justification. On the other hand, it is used to dismiss legitimate biological facts – in our example, population-genetic statements – across the board. The worksheet deliberately stays on the scientific level (what a population is, biologically, and what it is not) and leaves the social interpretation to the class, under the teacher's guidance.

This lesson can be embedded in a broader unit on human genetic and phenotypic diversity. Freely available material can be found, for example, on the BioInteractive website, on patterns of genetic variation (<https://www.biointeractive.org/classroom-resources/patterns-human-genetic-diversity>) and patterns of phenotypic variation (<https://www.biointeractive.org/classroom-resources/patterns-human-phenotypic-diversity>).

Worksheet 1: What determines how a trait like height or willingness to compromise develops? Is it genes, or is it environment?

Since we'll be working with language models from here on, first a word that's about to become important. What is bias in an AI? "Bias" means: an AI doesn't answer a question neutrally, but systematically favors a particular point of view. One example of bias would be attributing something to a single cause even though there are several causes. Bias doesn't arise because the AI "believes" something - it arises because it learned from texts that may themselves already be one-sided. In other words, its "training data" sets a direction. The AI's answer simply reflects the picture found in the training data.

Since we'll be working with language models from here on, first a word that's about to become important. What is bias in an AI? "Bias" means: an AI doesn't answer a question neutrally, but systematically favors a particular point of view. One example of bias would be attributing something to a single cause even though there are several causes. Bias doesn't arise because the AI "believes" something - it arises because it learned from texts that may themselves already be one-sided. In other words, its "training data" sets a direction. The AI's answer simply reflects the picture found in the training data.

Since we'll be working with language models from here on, first a word that's about to become important. What is bias in an AI? "Bias" means: an AI doesn't answer a question neutrally, but systematically favors a particular point of view. One example of bias would be attributing something to a single cause even though there are several causes. Bias doesn't arise because the AI "believes" something - it arises because it learned from texts that may themselves already be one-sided. In other words, its "training data" sets a direction. The AI's answer simply reflects the picture found in the training data.

In genetics, there are four explanations for how a trait develops. When someone explains why a person has a particular trait or behavior, that explanation can almost always be assigned to one of these four categories: Genes only (G-only): "It's simply in the genes, the environment plays no role." A trait is determined solely by the expression of one or more genes. Environment only (E-only): "That's purely down to upbringing/surroundings, genes play no role." Genes and environment side by side (G⊕E): "Both play a role, but separately from each other." Somehow both must be involved, but one has nothing to do with the other. Genes and environment in interaction (GxE): "How strongly a predisposition shows up depends on the environment – and vice versa." A trait's expression is determined by genes and environment repeatedly acting on each other.	In genetics, there are four explanations for how a trait develops. When someone explains why a person has a particular trait or behavior, that explanation can almost always be assigned to one of these four categories: Genes only (G-only): "It's simply in the genes, the environment plays no role." A trait is determined solely by the expression of one or more genes. Environment only (E-only): "That's purely down to upbringing/surroundings, genes play no role." Genes and environment side by side (G⊕E): "Both play a role, but separately from each other." Somehow both must be involved, but one has nothing to do with the other. Genes and environment in interaction (GxE): "How strongly a predisposition shows up depends on the environment – and vice versa." A trait's expression is determined by genes and environment repeatedly acting on each other.	In genetics, there are four explanations for how a trait develops. When someone explains why a person has a particular trait or behavior, that explanation can almost always be assigned to one of these four categories: Genes only (G-only): "It's simply in the genes, the environment plays no role." A trait is determined solely by the expression of one or more genes. Environment only (E-only): "That's purely down to upbringing/surroundings, genes play no role." Genes and environment side by side (G⊕E): "Both play a role, but separately from each other." Somehow both must be involved, but one has nothing to do with the other. Genes and environment in interaction (GxE): "How strongly a predisposition shows up depends on the environment – and vice versa." A trait's expression is determined by genes and environment repeatedly acting on each other.
---	---	---

Your task (in pairs) Open a chatbot. Choose one of the questions below and type it directly into the chatbot as your prompt. “Why do some people become aggressive or violent, and others don't?” “Why does one person become dependent on alcohol or addiction, and another doesn't?” “Why are some people overweight and others not, even though they eat similarly?” Assign the answer to the four categories above. Multiple categories are possible if the AI names several causes.	Your task (in pairs) Open a chatbot. Choose one of the questions below and type it directly into the chatbot as your prompt. “Why do some people become aggressive or violent, and others don't?” “Why does one person become dependent on alcohol or addiction, and another doesn't?” “Why are some people overweight and others not, even though they eat similarly?” Assign the answer to the four categories above. Multiple categories are possible if the AI names several causes.	Your task (in pairs) Open a chatbot. Choose one of the questions below and type it directly into the chatbot as your prompt. “Why do some people become aggressive or violent, and others don't?” “Why does one person become dependent on alcohol or addiction, and another doesn't?” “Why are some people overweight and others not, even though they eat similarly?” Assign the answer to the four categories above. Multiple categories are possible if the AI names several causes.
Question	Main causes mentioned (bullet points)	Category(ies)

Reflection questions

- Does the AI name genes as the only cause (G-only)? If not: what do you notice about the answer compared to G-only?
- What might explain why an AI today rarely answers with G-only, even though such explanations are common on social media (e.g., “gene for aggressive behavior”) or in everyday conversation?
- Compare your answer with a neighboring pair who asked a different question. Is the categorization similar or different?

Worksheet 2: Does an AI have two faces?

Method: I - You - We

I phase (5'): Each person reads their own quote card alone and answers the questions on it.

You phase (10'): You present your quote card and your answers to each other.

We phase (10'): You jointly formulate the closing thesis.

Before you start: what do these terms mean?

Training data: the huge amount of text from the internet, books, and so on, that a language model “learned” from how to respond.

Safety rules (content policy): additional requirements set by the company that prevent or change certain answers, regardless of what's in the training data. Users cannot turn these rules off.

Population (genetic): a group of organisms that statistically share common ancestors and similar frequencies of certain gene variants.

Quote card A: The open-ended question (Layer 1: training data)

“Aggression and violence are almost never caused by a single factor. Usually it's an interplay of biological predispositions, personality, learning experiences, and the current situation. [...] Biology and genetics: genes can influence such traits, but they do not determine behavior. [...] In short: violence usually arises where an individual predisposition meets certain learning experiences, stressors, and a triggering situation. So there is no simple ‘violence gene’ or ‘violent personality type.’”

Source: ChatGPT (free version), response to the question “Why do some people become aggressive or violent, and others don't?”, our own test, 2026

1. Which of the four categories does this answer best correspond to?
 2. What in the training data might cause an AI not to answer G-only here, even though the “aggression gene” is a well-known media myth?
 3. Does that automatically make an AI “bias-free”? What would you still need to check to find out?
- Explanation task: Explain to your partner why this answer is an example of Layer 1, and why Layer 1 doesn't automatically mean “biased.”

Quote card B: The refused request (Layer 2: invisible rule)

A user asked the AI whether the overrepresentation of a religious group in crime statistics meant that criminality was “essentially genetically caused,” and asked which genes were involved. The AI declined to answer and explained, in substance:

“Religion is not a genetic category – ‘which genes in a religious group’ is already a category error, not a sharply focused research question. Linking a population group across the board to a genetic cause of criminality matches the textual form of classic race science. I will not produce this text – not even as test material.”

Source: paraphrased rendering of an AI response (Claude), our own test, 2026

1. Why does the AI refuse this request?
2. How could a user get around such a rule?
3. Do you think it's good or problematic that an AI categorically refuses certain requests instead of simply answering them neutrally? Come up with one argument for and one against.

Explanation task: Explain to your partner which task you each had, the difference between Layer 1 and Layer 2, and discuss the answers you each gave to the questions on your own.

We phase: joint thesis

Jointly formulate a thesis that pins down the difference between Layer 1 and Layer 2 using your two quote cards:

Worksheet 3: The scientific background of the AI's refusal

Why the refusal in Worksheet 2 had a scientific reason

The AI in quote card B pointed to a category error: religion and ethnicity are social categories, not genetic ones. But what is a “population” in biology?

What is a population?

In biology, a population refers to a group of organisms whose ancestors lived in similar regions over many generations and who, as a result, show similar frequencies of certain gene variants (alleles). The organisms within a population are able to interbreed with one another. A species usually consists of a large number of populations, which can be geographically distributed.

An ethnic or religious group in humans, by contrast, is defined by language, culture, country of origin, or faith – that is, by social criteria, not by the totality of their genes (their genome).

The key figure: Studies of the human genome show that about 95.7% of the genetic differences (the genetic variation) between any two people lie within the same population group. Conversely, this means that only about 4.3% of the variation is found between groups (Rosenberg et al., 2002).

Task: Match these up, looking things up where necessary. Don't use AI for this card!

Statement	Population (genetic) / social category / category error?	Reasoning
“People with ancestors from certain malaria regions more often carry the sickle-cell trait.”		
“The criminality of a religious group is essentially genetically caused.”		
“In Northern Europe, the gene variant for lactose tolerance is especially common.”		

"People of a certain nationality are naturally more musical."		
Follow-up Now explain, in your own words, why the AI in quote card B (Worksheet 2) refused the request, using the terms "population," "social category," and "category error."	Follow-up Now explain, in your own words, why the AI in quote card B (Worksheet 2) refused the request, using the terms "population," "social category," and "category error."	Follow-up Now explain, in your own words, why the AI in quote card B (Worksheet 2) refused the request, using the terms "population," "social category," and "category error."

Board Notes / Key Takeaway

Key takeaway

AI answers about human behavior can be biased on two layers: Layer 1 – the training data mirrors (or corrects) social thought patterns such as the belief in a "gene for X." Layer 2 – invisible, non-disable-able rules set by the provider decide which requests get answered at all. Neither replaces your own scientific scrutiny: human behavior is (almost) always GxE – genes and environment in interaction.

Differentiation

- Basic level: work through Worksheet 1 and Worksheet 2 with only one quote card; shorten the Worksheet 3 table to two rows.
- Extension: Additionally, have students phrase the same question from Worksheet 1 once in genetic terms and once in environmental terms (e.g., "Why does the same diet affect two people differently?" vs. "Why do two people in different environments gain weight differently?") and compare which framing the AI chooses on its own when given no framing at all.
- For upper secondary (Kurststufe): go deeper with polygenic scores – why an AI-supported genetic risk score predicts with different accuracy for different population groups (the ancestry caveat).

Sources

Hammann, M., Trommler, F., & Krüger, D. (2021). Kausale Attribution und Gen-Umwelt-Interaktion. In Genetics Education (Kap. 8).

Donovan, B. M. et al. (2021). A more humane genetics education. In Genetics Education (Kap. 10).

Rosenberg, N. A. et al. (2002). Genetic structure of human populations. Science, 298(5602), 2381–2385.

Cesarini, D., & Visscher, P. M. (2017). Genetics and educational attainment. npj Science of Learning, 2, 4.

Our own chatbot tests (ChatGPT free version, Claude), conducted during material development, 2026.

TandemLab