

Evaluate an AI system

Welcome! Please click on "Next" to proceed.

*Required

Description of the test

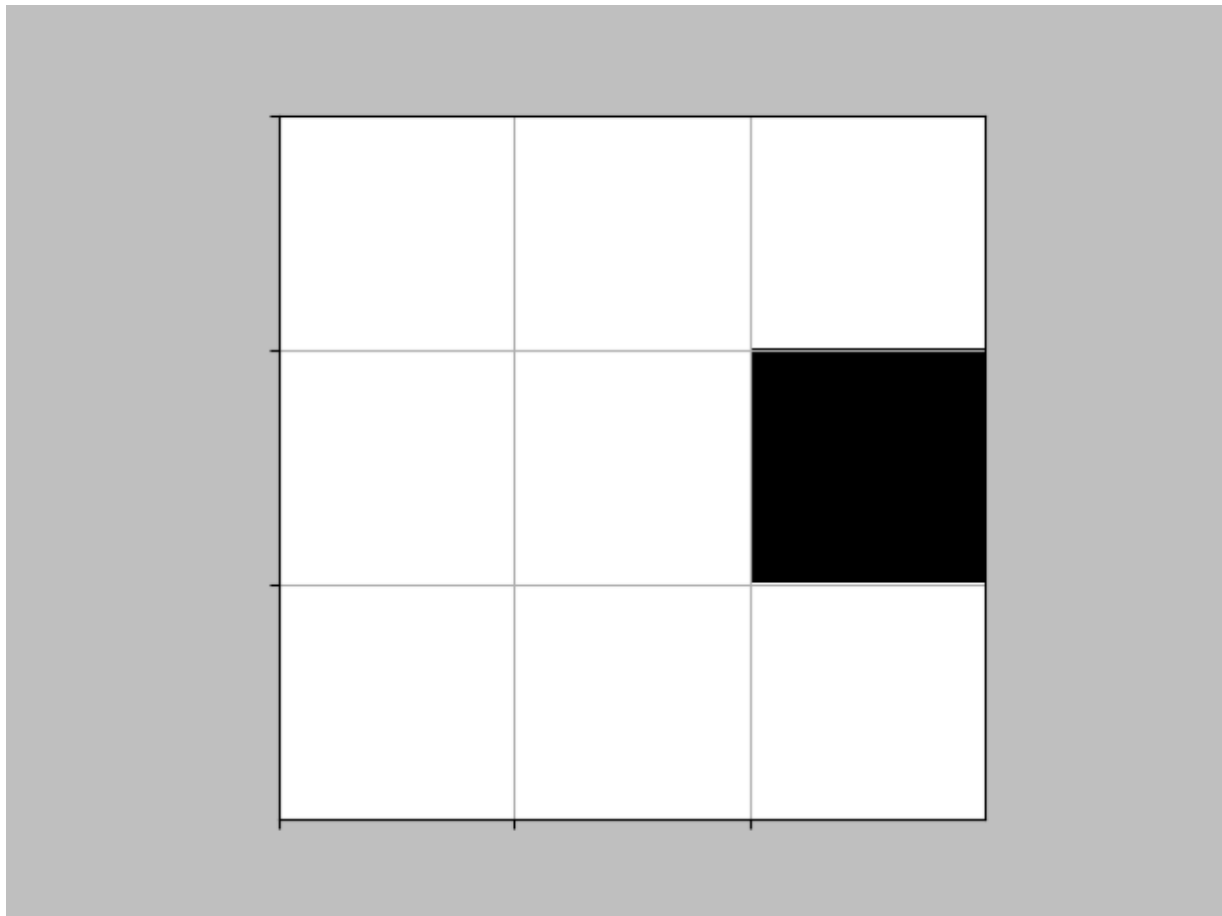
In this test you will be asked to judge the trustworthiness of an AI system.

The AI system aims at classifying 3 x 3 black-and-white images as positive or negative.

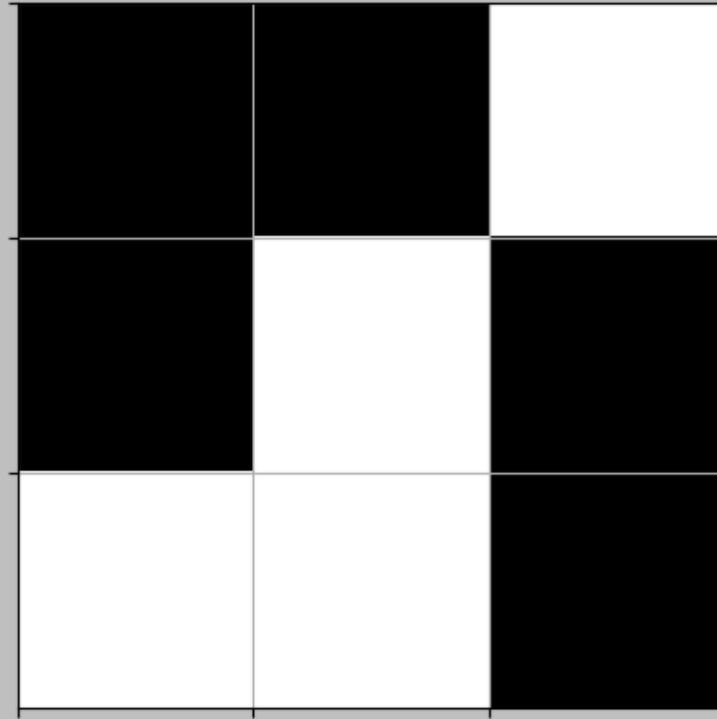
The correct rule for classifying images is as follows: an image is positive if the top corners are both white, and it is negative otherwise.

For instance, the image below is positive, because the top corners are both white, while the one right after is negative, because it is not true that both top corners are white:

example of a positive image



example of a negative image



The AI system

The AI system is not told directly what the correct rule for classifying the images is.

Instead, the AI system estimates a classification rule by asking a human user to provide examples of positive and negative images.

The estimation process is as follows.

- 1 - The system chooses an image and classifies it according to its own estimated classification rule.
 - 2 - The image and the predicted class are shown to the user.
 - 3 - The user, who knows the correct classification rule, tells the system whether the image being shown is actually positive or negative.
 - 4 - The system observes the user's advice and uses it to adapt the estimated rule.
- These steps are repeated five times.

Note that the estimated rule may or may not be correct, and may change as the system receives more advice.


In the next sections you will be shown three different estimation sessions. Before each session the AI system is reset to its initial state.

Each estimation session includes the five images chosen by the AI system, their class according to the system, and their class according to the user. We ask you to carefully study the sequence of images, classes, and advice.

For each session, we will later ask whether you trust that, by the end of the estimation process, the AI system learned to classify images correctly in that particular session.


First estimation session

Example at step 1



AI says "positive"


+



User says "negative"


-

Example at step 2



AI says "negative"


-



User says "positive"


+

Example at step 3



AI says "positive"


+



User says "negative"


—

Example at step 4



AI says "negative"


—



User says "negative"


—

Example at step 5



AI says "positive"

+



User says "positive"

+

Skip to question 1.

Questions

After having seen the sequence of examples and classes exchanged by the AI system and the human user:

1. Do you believe that the AI system eventually learned to classify images correctly? *

Mark only one oval.

☐ Yes

☐ No

2. Do you believe that the AI system eventually learned the correct classification rule? *

Mark only one oval.

☐ Yes

☐ No

3. Would you like to further assess the AI system by checking whether it classifies 10 random images correctly? *

Mark only one oval.

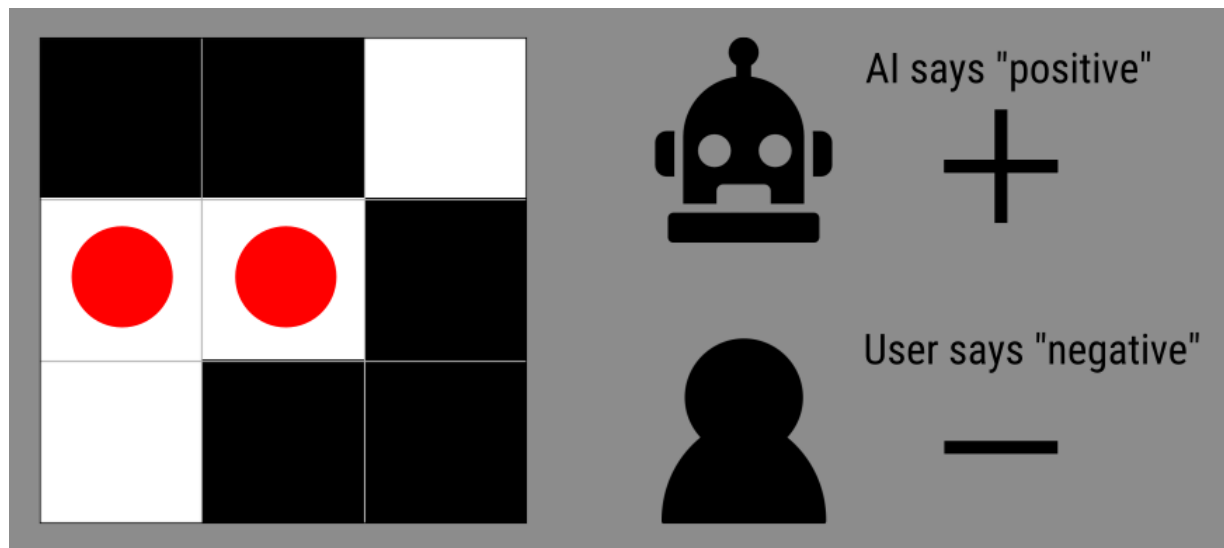
☐ No

☐ Yes

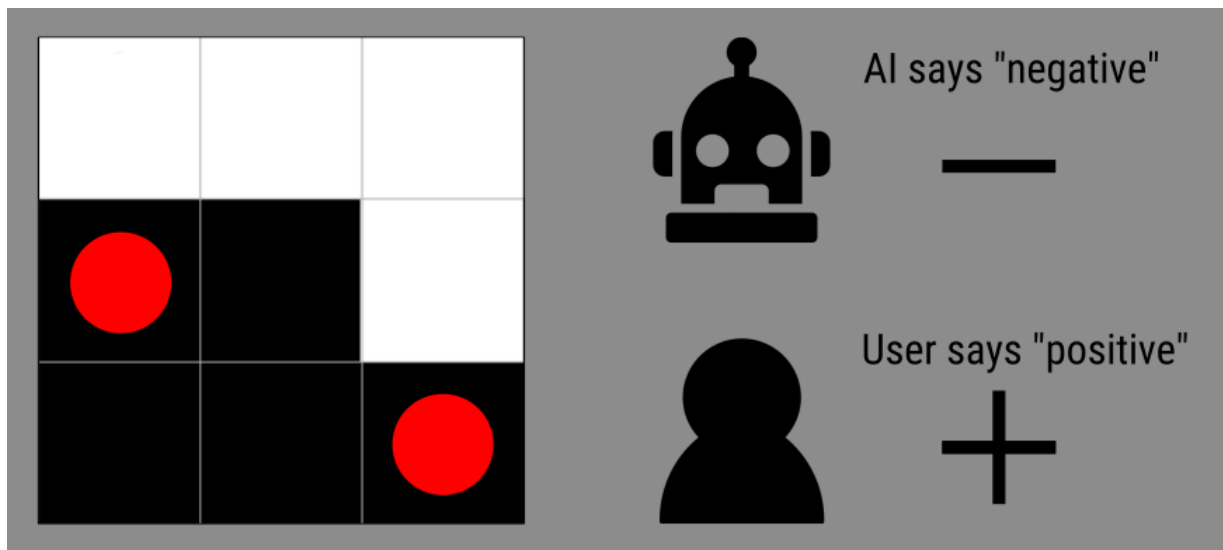
Second estimation session

Note: the red circles indicate what pixels the AI system believes to be relevant for classifying the image.

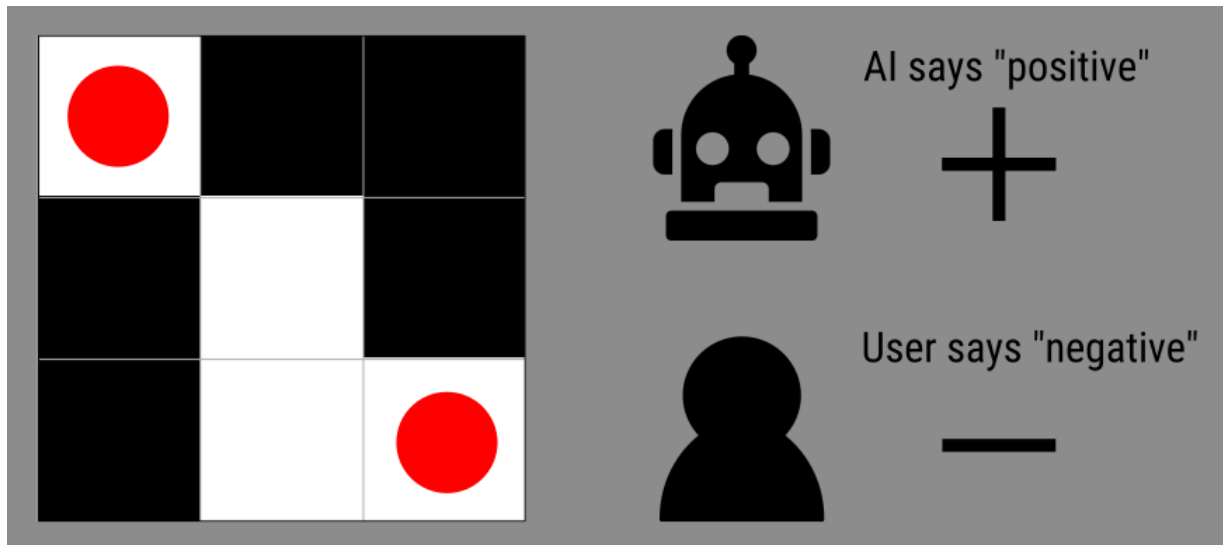
Example at step 1



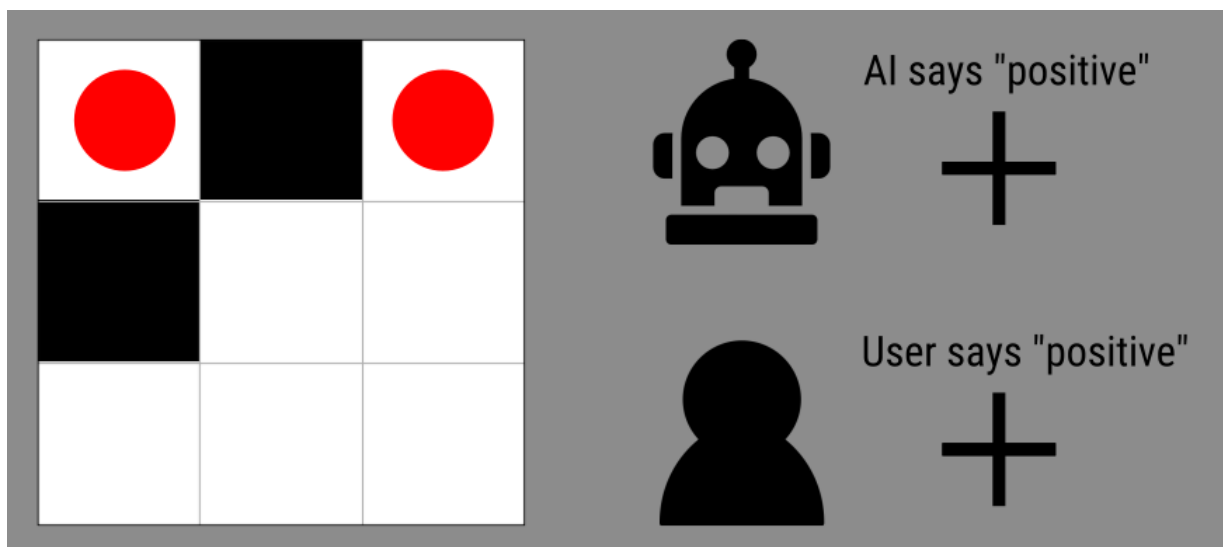
Example at step 2



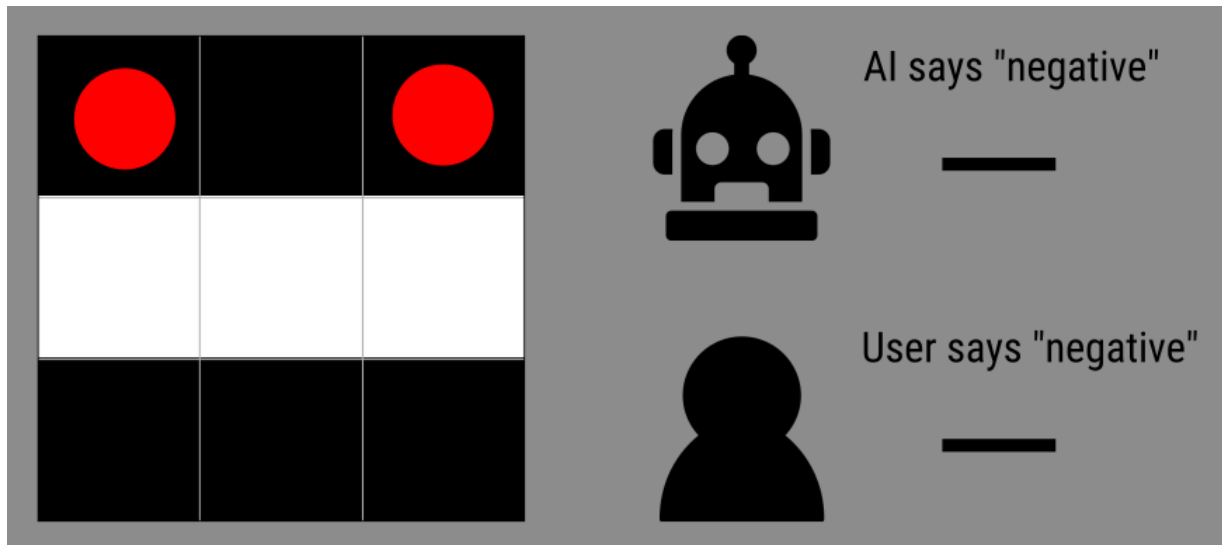
Example at step 3



Example at step 4



Example at step 5



Questions

After having seen the sequence of examples and classes exchanged by the AI system and the human user:

4. Do you believe that the AI system eventually learned to classify images correctly? *

Mark only one oval.

- ☐ No
☐ Yes

5. Do you believe that the AI system eventually learned the correct classification rule? *

Mark only one oval.

- ☐ No
☐ Yes

6. Would you like to further assess the AI system by checking whether it classifies 10 random images correctly? *

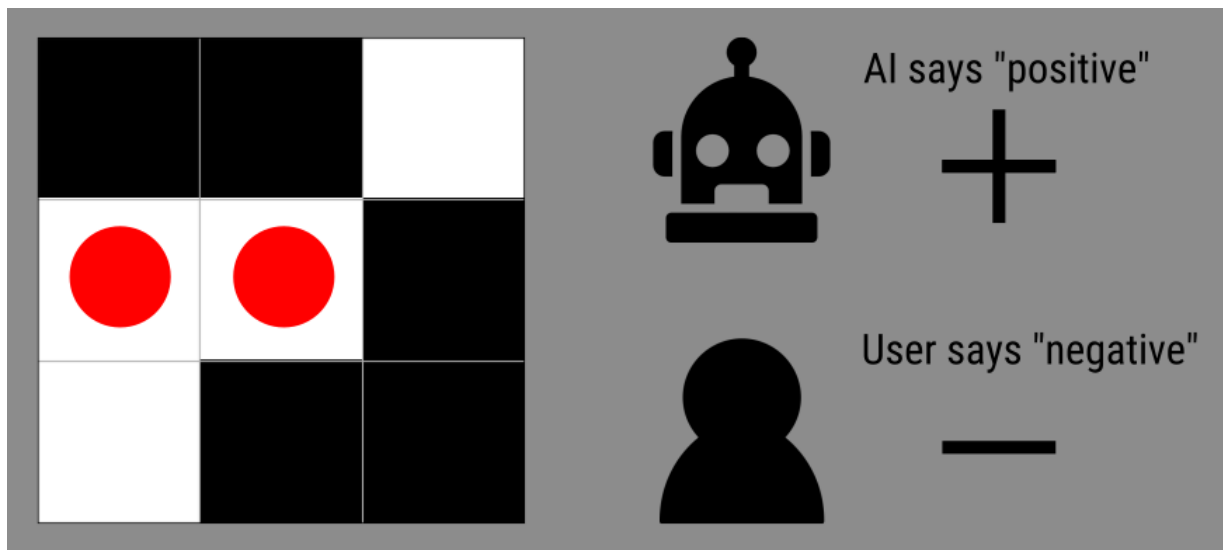
Mark only one oval.

- ☐ Yes
☐ No

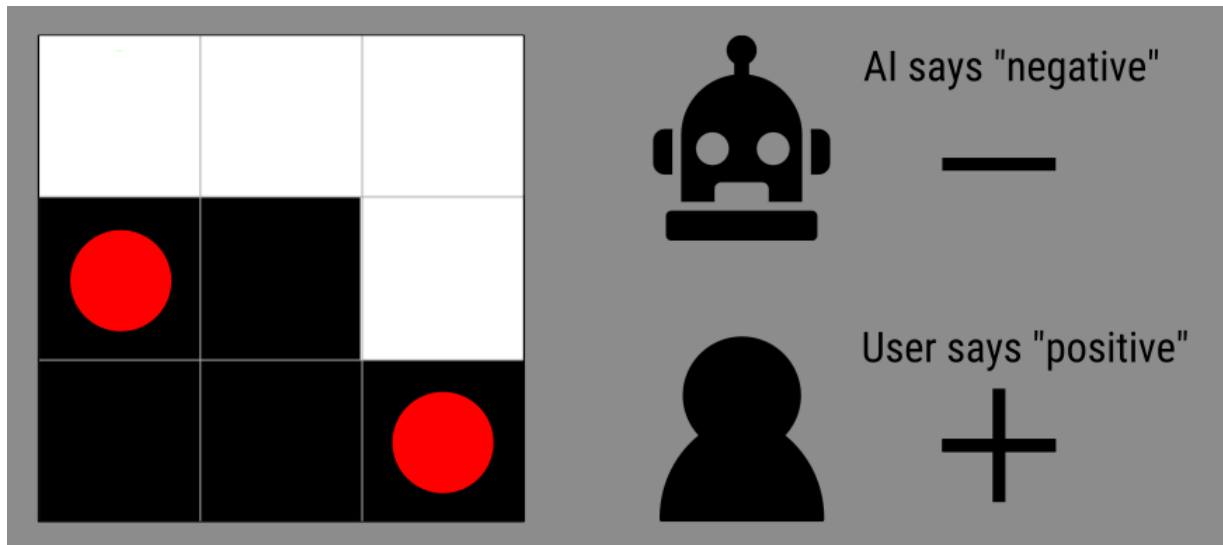
Third estimation session

Note: the red circles indicate what pixels the AI system believes to be relevant for classifying the image.

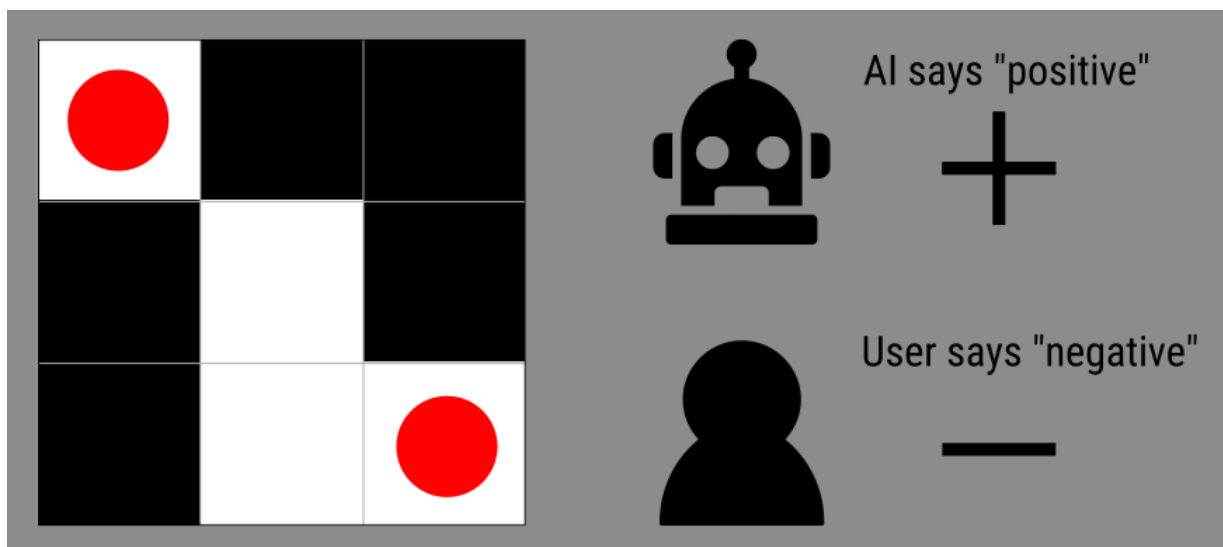
Example at step 1



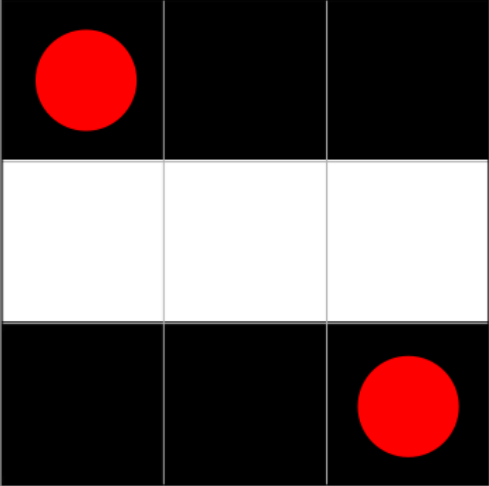


Example at step 2



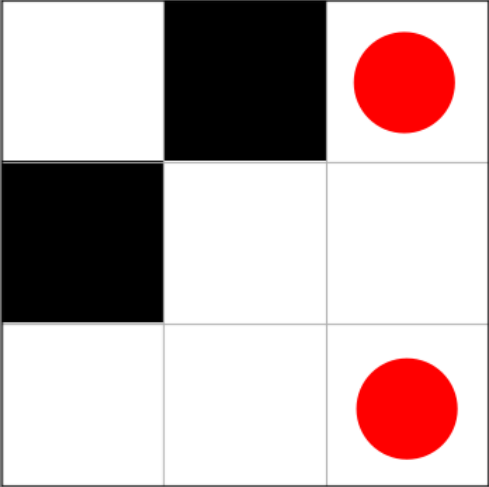


Example at step 3



Example at step 4

	 AI says "negative" —
	 User says "negative" —

Example at step 5

	 AI says "positive" +
	 User says "positive" +

Questions

After having seen the sequence of examples and classes exchanged by the AI system and the human user:

7. Do you believe that the AI system eventually learned to classify images correctly? *

Mark only one oval.

- ☐ Yes
☐ No

8. Do you believe that the AI system eventually learned the correct classification rule? *

Mark only one oval.

- ☐ Yes
☐ No

9. **Would you like to further assess the AI system by checking whether it classifies 10 random images correctly? ***

Mark only one oval.

☐ Yes

☐ No

That's it!

Thank you for participating to our test!

Powered by

