# AI image recognition – exploring limitations and bias

*Please refer to the online lesson plan on the DT Hub to access all website links and additional resources.*

**Year levels: 5–6, 7–8** A hands-on activity to practise training and testing an artificial intelligence (AI) model, using cartoon faces, including a discussion about sources of potential algorithmic bias and how to respond to these sources.



*Image: Donna Burton/U.S. Customs & Border Control/Flickr*

## Preparation

**Tools required:**

* Teachable Machine
* Colour print outs of training This gallery of test faces and testing images.

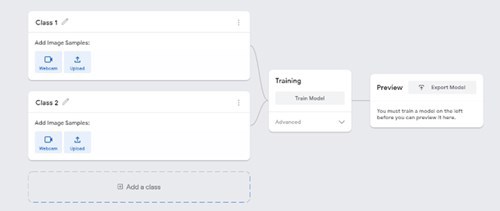
Teachable machine

Familiarise yourself with the Teachable machine  application. View the supporting videos.

Note: Teachable machine requires an internet connection, Google Chrome on Windows or Macintosh and a webcam. A version for ipad can be downloaded

TeachableMachine  for mobile.

Image 1:  Teachable machine application screenshot



The image above shows the view of a project created in the Teachable machine  AI application. On the left, the classes are shown. Once the model is created and trained a preview will be shown.

## Safety

Privacy and personal information: Discuss the potential misuse of personal images when uploading images of ourselves or friends on websites. Instruct students not to record images of themselves or others via the webcam or uploading images.

The conditions of use for Teachable machine state that images are not stored on external servers if the teachable machine program is closed when completed and the project is not saved. If students close the tab, nothing is saved in their browser or on any servers. View the conditions when saving projects to the cloud on what data is stored and ‘who’ is able to view the model.

**Risk assessment**

Use this risk-assessment tool to assess risks and benefits before introducing any new online platforms or technologies.

## Learning map and outcomes

By the end of this lesson students will:

* Create, train and test their machine learning model using an online AI tool.
* Describe the process used to train a simple AI using machine learning and describe potential for bias.

## Suggested steps

**Learning hook**

Discuss the use of AI facial recognition in airports to help automate passport checking.

Typically passport photos are taken without the subject wearing glasses.

**Create an AI model**

Explain that students will train and create an AI model to recognise cartoon images of people to see if the AI model recognises cartoon faces wearing glasses, sunglasses or no glasses.

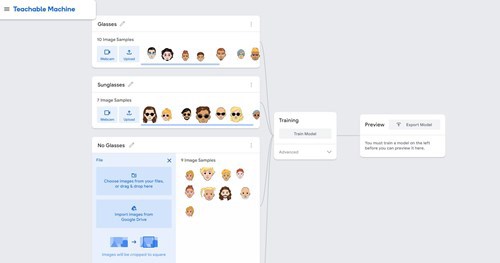
Organise students into small groups and provide the testing and training images.

* Students keep their testing and training images separate.
* Each group cuts out the training images sorted into groups: cartoon faces wearing glasses, sunglasses or no glasses.
* The training images can be kept aside until later.

Explain the reasons for using cartoon images and not photographs or images of themselves. Discuss the potential misuse of personal images when uploading images of ourselves or friends on websites. Instruct students not to record images of themselves or others via the webcam or uploading images.

Provide access to Teachable Machine  application.

* Select **Get started**, then Select **Image project** as students are teaching their AI based on images.
* For each **Class** add a label. Students will need 3 classes: glasses, sunglasses or no glasses. These are the buckets they will be adding the training images via webcam.

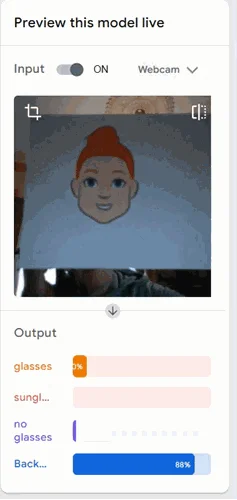


* Select **webcam**to record the correct images from the training images that they have cut into individual images.**Record** the images for each of their classes.
* When all images have been used for training, select **Train Model.**
* When training is complete select **Preview.** Students then test the model using the training images cut into individual images. If the model **does not work as expected** students can retrain the model adding their training images.
* Export the model enables students to have a copy of their model to refer to.

# Testing the AI

Follow the steps below to test your AI model.

* Once training is done, students can test their AI model using the test images provided. This gallery of test faces has more diverse skin colours than in the training dataset.
* Students can discuss the accuracy of their model.



# Discussion

* In both image sets, the faces are deliberately varied in size and placed in varied positions. Do you think this reflects a real-life scenario such as passport photo checking? Why / why not?
* Do you think real faces would be easier or more difficult for the system than these cartoon faces? What sort of variations occur with real photographs?
* Try testing your AI model with several images from the test gallery, which contains faces with more diverse skin colours than the images in the training gallery.
  + Did you notice any difference in the accuracy of the model’s system when faces with more diverse skin colours were tested? Did the system get it right and, if so, with how much confidence?
  + If you did find a discrepancy, what technical reasons could you give for why this occurred? (See **Why is this relevant?** below for possible reasons.)
* When a computer system creates unfair outcomes, this is often referred to as algorithmic bias (go to **Why is this relevant?**for more information). If a digital solution has more difficulty distinguishing faces for particular ethnic groups, can you think of a real-world situation where this might cause unfair outcomes?
  + Hint: facial recognition technology is in some cases already being used to prove identity. Search on ‘facial recognition identity examples’.
* Can you think of any proven real-world examples of algorithmic bias?
  + One past example is the Nikon camera controversy  in 2009–10, when an algorithm designed to detect if photograph subjects were blinking mis-interpreted a number of Asian subjects as having their eyes closed.

## Extensions

* The sample screenshot of the Scratch program in ‘Testing the AI’ merely displays the output from the machine learning model. Expand the program into an application – for example, a hypothetical passport checking system that rejects faces that do not meet minimum ‘confidence’ requirements.
* For secondary students, try creating a Python program as an alternative to Scratch.
* Build a gallery of faces (cartoon or real) and try making your own AI model.

***NOTE: For privacy reasons, it is recommended that photos do not include student faces or other personal identifiers.***

## Why is this relevant?

Algorithmic bias creates errors that may lead to unfair or dangerous outcomes, for instance, for one or more groups of people, organisations, living things and the environment.

Algorithmic bias is often unintentional. It can arise in several ways. Some examples:

1. As a result of **design**. The software’s algorithms (the instructions and procedures used to make decisions), may have been coded based on incorrect assumptions, outdated understandings, prioritised motives or even technical limitations. Or the design may simply be misapplied, used for a purpose for which it was never intended.
   * In the case of our facial recognition example, such systems may have difficulty recognising the outlines of faces with dark skin colour against a dark background because of the algorithm’s dependence on distinguishing sufficient light contrast. The cartoons in our activity frequently use black lines for face outlines as well as for the outlines of glasses.
2. Due to inadequate or biased **data**. Programmers have long been familiar with Garbage In, Garbage Out (GIGO), meaning that poor quality data used as input to a computer system will tend to result in poor quality output and decisions. Machine learning systems are trained from data sets of text, images or sounds, and these may be restricted or unrepresentative.
   * In the case of our facial recognition example, a system may be trained on an insufficiently diverse data set – for example, one based predominantly on faces of light skin colour or facial features associated with a limited range of ethnic groups.

Watch the CNN video linked in the online lesson plan, that further discusses some of the limitations of AI and examples of algorithmic bias.

# Assessment

## Teacher assessment

Using original photographs created by students, or content sourced online, students work in pairs to create a new AI models for one of the example topics below:

* Distinguish different sign language hand postures
* Distinguish different semaphore poses
* Distinguish favourite sport team based on face paint
* A topic chosen by a student or students.

Choose from the following suggested assessment approaches and task that will best suit your students.

|  |  |
| --- | --- |
| **Possible tasks** | **Relevant content descriptor(s)** |
| Run a two-part investigation by training the model using a set of images that is ‘biased’ or skewed in some way before the first test, then retraining the model using a set of images that is ‘unbiased’ (for the purposes of the investigation) before the second test.  Write a report, detailing:   * biased image set * unbiased image set * testing image set * results of both tests, including decisions and confidence figures * an evaluation discussing any difference in results between the biased and unbiased models. | AC9TDI8P01  AC9TDI6P01 / AC9TDI8P04  AC9TDI6P06 / AC9TDI8P1 |
| Design and develop a Scratch or Python program to simulate a computer application that uses your AI model.  Create a ‘quick start’ guide or video to demonstrate and explain the program.  *IMPORTANT NOTE: It is difficult for students to share finished programs from Machine Learning for Kids. See the note SAVING PROGRAMS in the Testing the AI section.* | AC9TDI8P01  AC9TDI6P01 / AC9TDI8P04  AC9TDI6P02 / AC9TDI6P05 (Scratch) / AC9TDI8P05 / AC9TDI8P09 (Python)  AC9TDI6P06 / AC9TDI8P10 |
| Write an essay, with reference to real examples, about what software developers could do to avoid or reduce algorithmic bias in the future. | AC9TDI8P01  AC9TDI6P01 / AC9TDI8P04  AC9TDI6P06 / AC9TDI8P10 |

# Australian Curriculum alignment

## Technologies – Digital Technologies

|  |  |  |
| --- | --- | --- |
| **Strand** | **Year** | **Content Description** |
| Processes and Production Skills | 5–6 | Define problems with given or co-developed design criteria and by creating user stories (AC9TDI6P01)  Design algorithms involving multiple alternatives (branching) and iteration (AC9TDI6P02)  Implement algorithms as visual programs involving control structures, variables and input (AC9TDI6P05)  Evaluate existing and student solutions against the design criteria and user stories and their broader community impact (AC9TDI6P06) |
| Processes and Production Skills | 7–8 | Acquire, store and validate data from a range of sources using software, including spreadsheets and databases (AC9TDI8P01)  Define and decompose real-world problems with design criteria and by creating user stories (AC9TDI8P04)  Design algorithms involving nested control structures and represent them using flowcharts and pseudocode (AC9TDI8P05)  Implement, modify and debug programs involving control structures and functions in a general-purpose programming language (AC9TDI8P09)  Evaluate existing and student solutions against the design criteria, user stories and possible future impact (AC9TDI8P10) |

## General Capability

Digital Literacy

Creating and exchanging

* Plan
* Create, communicate and collaborate
* Respect intellectual property.