“Breaking the AI” - Technical Exercise
for
Microsoft EPIC “Artificial Intelligence and Engineering Design Process” Session

Learning Objectives:

- Gain hands-on understanding of how an AI system functions under real-world conditions
- Practice critical thinking and technical troubleshooting
- Apply understanding of AI to identify constraints and improvement opportunities with an existing technology
- Have FUN!

During this session, students will practice technical skills relating to AI by attempting to break a strong AI system – the Microsoft Surface Facial Recognition app.

Background:

The facial recognition app on the surface is capable of distinguishing between one human face and another, based on the distinctiveness of certain facial characteristics. First the user must take a picture of his/her face to serve as a reference, and then the AI should (in theory) be able to determine with a strong degree of accuracy whether future images are a match to that individual.

The app is also able to make general determinations including age, gender, and mood, based on a pre-existing database of reference images for human facial characteristics.

However, as with any computer system, the AI is subject to error and can make mistakes. The accuracy of the AI depends on the quality and quantity of data with which it was trained, as well as the quality of the user defined data (captured images).

Systems used in law enforcement, crowd control, security and risk assessment, and many other current applications function under the same basic constraints of the Surface Facial Recognition app.

By understanding how this consumer level AI work (and more importantly WHY it often fails), students will gain practical insight into the current state and reliability of AI technology, as well as the opportunities for improving these systems in the future.

"Break the AI" Exercise

Stage 1: Functioning of AI under optimal conditions:

For the first 10min, students are to use the Surface Facial Recognition app as intended:

- Take a reference image of your face
- Make sure that your face is not obscured (remove glasses, ensure proper lighting, avoid blocking face with hair, etc.)
- Change your expression slightly (do not contort face) and move to a different location; take another image and allow AI to compare
- Find pictures of your face on your phone or other source and use the Surface to take a picture of that image; test the system to see if it can properly identify you through a secondary image
- Note the overall accuracy of the identification, gender, age, and emotion assessments of the AI

Stage 2: Functioning of AI under adverse conditions:

After establishing a baseline, students will attempt to disable the AI using increasingly extreme measures to alter their appearance.

- Once you have tested the AI under optimal/normal conditions, you will no attempt to "break" or trick the AI

- First, try to trick the AI by making minor changes to your hair and slightly more extreme facial expressions
- Next, use clothing to alter your appearance, while leaving your face properly exposed (put on a hoodie to cover hair, wear a headband, put on normal prescription style glasses, etc.)
- Finally, try extreme measures to trick the AI
    - Obscure eyes with sunglasses
    - Put on a medical face mask that covers mouth
    - Use props to more significantly alter your appearance – wigs, makeup, feathers, face paint, fake mustache/beard, etc.
    - Test all aspects of the AI to determine which are most affected by your attempts to confuse or disable the facial recognition

Debrief (the most important part):

After student has completed the Stage 1 and Stage 2 experiments, the facilitator (parent/teacher) should lead a debrief, asking key questions to connect the exercise to what the student has already learned about AI.

Sample prompts/questions:

- How accurate was the AI overall?
- In specific areas (identity, gender, age, emotion)?
- If it made mistakes, why do you think it had trouble?
- For instance, why might the AI have trouble determining gender in younger students (10-16yrs/old)?
- How did the AI respond to the minor adjustments you made during Stage 1? Why?
- How did the AI respond to the more extreme adjustments you made during Stage 2? Why?
- Were there any particular adjustments that had a more significant impact than others? (e.g. covering hair, obscuring eyes, particular expressions, etc.)
- Do you notice any similarities between the types of things that disabled or tricked the AI and what we see in the real world when people want to hide their identities (e.g. celebrities and criminals hiding their eyes, changing or covering their hair, etc.)?
- Based on your experience, would you be comfortable with an AI being used to identify people as suspects in criminal cases? Or to assess an individual's emotional state in various situations (e.g. airport security, crowd control, etc.). Why?
- What could engineers do to improve the reliability of AI systems in the future?
- Extra credit: Why do you think the AI may have trouble properly determining age? Hint – Much of the information used to train the AI comes from social media posts and images.

- Why might it be dangerous if people don’t realistically understand the limitations and capabilities of AI systems like this one? What could go wrong?
- Continue with this discussion and push students to critically consider the societal role of AI and how these emergent technologies may change our lives for better or worse.

Recommended supplemental sources to further discussion:

Terminator, Ghost in the Shell, Sword Art Online, Alita: Battle Angel, The Matrix, Ex Machina, I, Robot.