7 Revealing Ways AIs Fail

Based on C. Q. Choi (2021). 7 Revealing Ways AIs Fail, IEEE Spectrum, October 2021, pp. 42-47.

https://spectrum.ieee.org/ai-failures

Brittleness

- Take a picture of a school bus. Flip it so it lays on its side, as it might be found in the case of an accident in the real world.
  - A 2018 study found that state-of-the-art AIs that would normally correctly identify the school bus right-side-up failed to do so on average 97 percent of the time when it was rotated.
  - Such a failure is an example of brittleness. An AI often "can only recognize a pattern it has seen before," Nguyen says. "If you show it a new pattern, it is easily fooled."
- Examples:
  - Fastening stickers on a stop sign can make an AI misread it.
  - Changing a single pixel on an image can make an AI think a horse is a frog.
  - Neural networks can be 99.99 percent confident that multicolor static is a picture of a lion.
  - Medical images can get modified in a way imperceptible to the human eye so medical scans misdiagnose cancer 100 percent of the time.
- One possible way to make AIs more robust against such failures is to expose them to as many confounding "adversarial" examples as possible, Hendrycks says.
  - However, they may still fail against rare " black swan" events. "Black-swan problems such as COVID or the recession are hard for even humans to address—they may not be problems just specific to machine learning," he notes.

Soh's Notes: Lack of transfer learning, lack of abstraction and generalization, lack of contextual understanding

Embedded Bias 1

- Increasingly, AI is used to help support major decisions, such as who receives a loan, the length of a jail sentence, and who gets health care first.
- The hope is that AIs can make decisions more impartially than people often have, but much research has found that biases embedded in the data on which these AIs are trained can result in automated discrimination en masse, posing immense risks to society.
- Case Study:
  - In 2019, scientists found a nationally deployed health care algorithm in the United States was racially biased, affecting millions of Americans. The AI was designed to identify which patients would benefit most from intensive-care programs, but it routinely enrolled healthier white patients into such programs ahead of black patients who were sicker.
  - Physician and researcher Ziad Obermeyer at the University of California, Berkeley, and his colleagues
    - found the algorithm mistakenly assumed that people with high health care costs were also the sickest patients and most in need of care
    - However, due to systemic racism, "black patients are less likely to get health care when they need it, so are less likely to generate costs," he explains.

Soh's Notes: Inadequate training (data samples not appropriate); lack of proper variables; lack of explanation & verification

Embedded Bias 2

- Case Study, Continued …
  - After working with the software's developer, Obermeyer and his colleagues helped design a new algorithm that analyzed other variables and displayed 84 percent less bias.
    - "It's a lot more work, but accounting for bias is not at all impossible," he says.
  - They recently drafted a playbook that outlines a few basic steps that governments, businesses, and other groups can implement to detect and prevent bias in existing and future software they use
    - Identifying all the algorithms they employ
    - Understanding this software's ideal target and its performance toward that goal
    - Retraining the AI if needed, and
    - Creating a high-level oversight body.

Soh's Notes: Inadequate training (data samples not appropriate); lack of proper variables; lack of explanation & verification

Catastrophic Forgetting 1

- Deepfakes—highly realistic artificially generated fake images and videos, often of celebrities, politicians, and other public figures—are becoming increasingly common on the Internet and social media, and could wreak plenty of havoc by fraudulently depicting people saying or doing things that never really happened.
- To develop an AI that could detect deepfakes, computer scientist Shahroz Tariq and his colleagues at Sungkyunkwan University, in South Korea, created a website where people could upload images to check their authenticity.
  - In the beginning, the researchers trained their neural network to spot one kind of deepfake.
  - However, after a few months, many new types of deepfake emerged, and when they trained their AI to identify these new varieties of deepfake, it quickly forgot how to detect the old ones.
- This was an example of catastrophic forgetting—the tendency of an AI to entirely and abruptly forget information it previously knew after learning new information, essentially overwriting past knowledge with new knowledge. "Artificial neural networks have a terrible memory," Tariq says.

Soh's Notes: Lack of understanding of temporal and spatial contexts, lack of understanding of paradigm shifts and phase transitions

Catastrophic Forgetting 2

- AI researchers are pursuing a variety of strategies to prevent catastrophic forgetting so that neural networks can, as humans seem to do, continuously learn effortlessly.
  - A simple technique is to create a specialized neural network for each new task one wants performed—say, distinguishing cats from dogs or apples from oranges—"but this is obviously not scalable, as the number of networks increases linearly with the number of tasks," says machine-learning researcher Sam Kessler at the University of Oxford, in England.
  - One alternative Tariq and his colleagues explored as they trained their AI to spot new kinds of deepfakes was to supply it with a small amount of data on how it identified older types so it would not forget how to detect them. Essentially, this is like reviewing a summary of a textbook chapter before an exam, Tariq says.
    - Knowledge distillation – learning from other AIs

Soh's Notes: Lack of understanding of temporal and spatial contexts, lack of understanding of paradigm shifts and phase transitions

Explainability

- Why *does* an AI suspect a person might be a criminal or have cancer? The explanation for this and other high-stakes predictions can have many legal, medical, and other consequences.
  - The way in which AIs reach conclusions has long been considered a mysterious black box, leading to many attempts to devise ways to explain AIs' inner workings.
- Nguyen and his colleagues investigated seven different techniques that researchers have developed to attribute explanations for AI decisions
  - E.g., what makes an image of a matchstick a matchstick? Is it the flame or the wooden stick?
  - They discovered that many of these methods "are quite unstable," Nguyen says. "They can give you different explanations every time."
  - In addition, while one attribution method might work on one set of neural networks, "it might fail completely on another set," Nguyen adds.
- The future of explainability may involve building databases of correct explanations
  - Attribution methods can then go to such knowledge bases "and search for facts that might explain decisions," he says.

Soh's Notes: Lack of explainability, contextual modeling and understanding

Quantifying Uncertainty

- In 2016, a Tesla Model S car on autopilot collided with a truck that was turning left in front of it in northern Florida, killing its driver— the automated driving system's first reported fatality. According to Tesla's official blog, neither the autopilot system nor the driver "noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied."
  - One potential way Tesla, Uber, and other companies may avoid such disasters is for their cars to do a better job at calculating and dealing with uncertainty.
- Currently AIs "can be very certain even though they're very wrong," Oxford's Kessler says that if an algorithm makes a decision, "we should have a robust idea of how confident it is in that decision, especially for a medical diagnosis or a self-driving car, and if it's very uncertain, then a human can intervene and give [their] own verdict or assessment of the situation."
- E.g.,, computer scientist Moloud Abdar at Deakin University in Australia and his colleagues applied several different uncertainty quantification techniques as an AI classified skin-cancer images as malignant or benign, or melanoma or not
  - The researcher found these methods helped prevent the AI from making overconfident diagnoses.
- Autonomous vehicles remain challenging for uncertainty quantification, as current uncertainty-quantification techniques are often relatively time consuming, "and cars cannot wait for them," Abdar says. "We need to have much faster approaches."

Soh's Notes: Lack of reasoning with uncertainty, lack of scalability, lack of contextual modeling and understanding

Common Sense

- AIs lack common sense—the ability to reach acceptable, logical conclusions based on a vast context of everyday knowledge that people usually take for granted
    - "If you don't pay very much attention to what these models are actually learning, they can learn shortcuts that lead them to misbehave," says computer scientist Xiang Ren at the University of Southern California.
- E.g., Scientists may train AIs to detect hate speech on data where such speech is unusually high, but such systems do not recognize the context well: "humans reading through a whole sentence can recognize when an adjective is used in a hateful context."
- Previous research suggested that state-of-the-art AIs could draw logical inferences about the world with up to roughly 90% accuracy
    - However, when Ren and his colleagues tested these models, they found even the best AI could generate logically coherent sentences with slightly less than 32 percent accuracy.
    - When it comes to developing common sense, "one thing we care a lot [about] these days in the AI community is employing more comprehensive checklists to look at the behavior of models on multiple dimensions," he says.

Soh's Notes: Lack of reasoning with uncertainty (fusion), lack of explainability, lack of contextual modeling and understanding

Math

- AIs "are surprisingly not good at mathematics at all," Berkeley's Hendrycks says. "You might have the latest and greatest models that take hundreds of GPUs to train, and they're still just not as reliable as a pocket calculator."
- E.g., Hendrycks and his colleagues trained an AI on hundreds of thousands of math problems with step-by-step solutions.
  - However, when tested on 12,500 problems from high school math competitions, "it only got something like 5 percent accuracy," he says.
  - In comparison, a three-time International Mathematical Olympiad gold medalist attained 90 percent success on such problems "without a calculator," he adds.
- Neural networks nowadays can learn to solve nearly every kind of problem "if you just give it enough data and enough resources, but not math," Hendrycks says.
- It remains uncertain why AI is currently bad at math
  - One possibility is that neural networks attack problems in a highly parallel manner like human brains, whereas math problems typically require a long series of steps to solve, so maybe the way AIs process data is not as suitable for such tasks, "in the same way that humans generally can't do huge calculations in their head," Hendrycks says.

Soh's Notes: Lack of reasoning with uncertainty (fusion), lack of explainability, lack of contextual modeling and understanding