Risks of AI systems

Prof. Neil C. Rowe

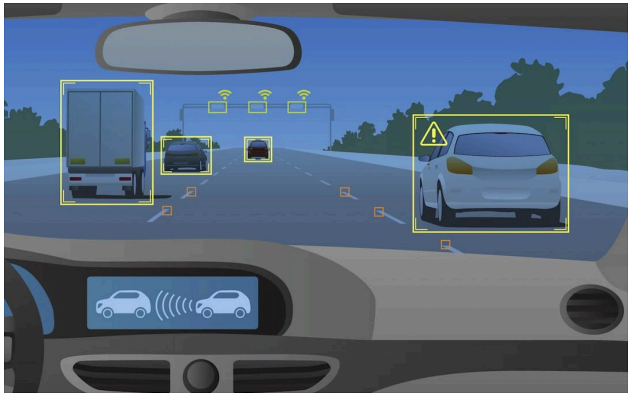
Computer Science Department
Naval Postgraduate School

http://faculty.nps.edu/ncrowe

Fall 2019

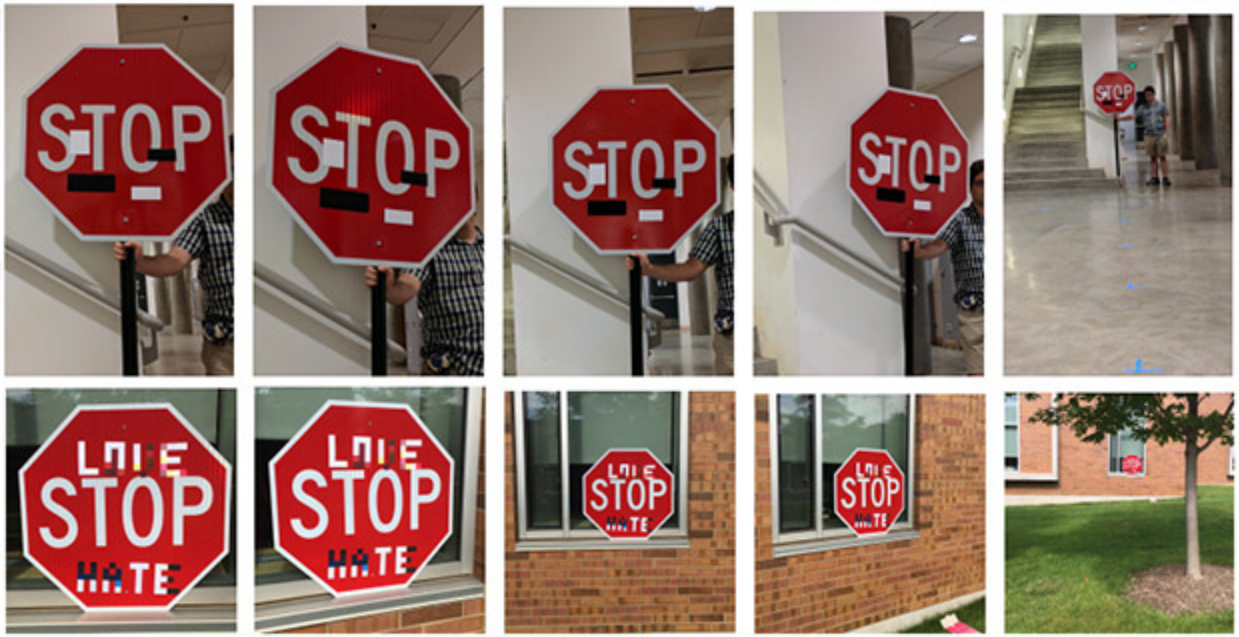
Risk: AI design can be oversimplified

- Driverless cars cannot plan for every rare road situation. So they may cause accidents.
- Similarly, automated machine guns (in Korea) and lethal drones (in the Middle East) cannot understand everything they see, e.g. attempts to surrender.
- Financial pressures may encourage vendors to oversimplify AI. Neural nets can be an oversimplification.
- Oversimplified AI may do bad things an ethical person would not.



The stop sign example

These were all misclassified as either speed limit 45 signs or yield signs. Stickers were affixed to a real stop sign after lengthy experimentation with a neural net trained to recognize stop signs to find the smallest modifications that would cause it to fail. The neural net is clearly not using the features that humans use to identify stop signs.



Risk: Bugs in AI can cause harms

- AI software has more bugs than other software because it's usually complex and probabilistic.
- Neural networks are especially complex.
- People can die or be hurt (including financially) because of faulty AI. Example: stop-sign misinterpreted as speed-limit sign by a robot car.
- Legal responsibility for faulty AI should reside with the writers of software. But proving responsibility can be difficult.
- Failure of some AI to explain itself exacerbates the problem.
- AI software can be hacked to emplace bugs.

Risk: Blame becomes harder with AI

- It is harder to see who to blame when things go wrong with AI systems than with other software.
- Unfairly blaming the AI may be easier than blaming bad human decisions behind it.
- This can increase the incompetence of organizations since they don’t get as much feedback about their failures.
- Good explanation capabilities in AI systems reduce blame problems. Rule-based systems explain better than neural networks.

Risk: Machine learning may disappoint

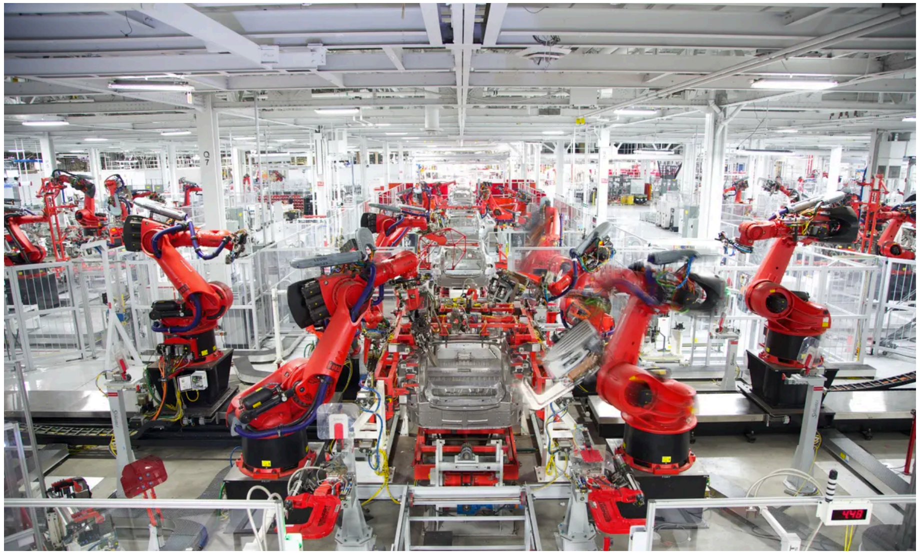
- Humans like to see patterns. In fact, they see patterns where none exist.
- Examples: gambling, conspiracy theories, supernatural phenomena.
- Machine learning is more accurate and honest in seeing patterns than humans are. Thus, it may disappoint humans eventually because it doesn't see all the patterns they see.

Risk: AI won’t know all human needs

- AI creates obedient servants, so it should be considering human needs as primary.
- However, it’s difficult to remember to tell the AI all the human needs to consider.
- For instance in an emergency, an AI needs to assign different priorities than normal, and not just continue routine activity.

Risk: AI is automation, and automation has potential harms

- AI creates unemployment of skilled (white-collar) labor, unlike older automation of unskilled labor.
- All unemployment creates bored people and social unrest.
- AI may cause a society where most people live on government support.



Risk: AI increases inequality in society

- Most technology increases inequality between those who have it and those who don't. Usually only temporarily, but it may take time to balance.
- Will everyone get AI, or only the rich and powerful?

- How long will it take the benefits of AI to disperse? Will they ever reach the poorer parts of Africa? Are they willing to wait?
- Rules for awarding mortgages could be based on business rather than moral considerations. A European law tries to prohibit this.

Risk: AI makes it easier to violate the privacy of people

- Suspicious people can be automatically classified based on a few clues.
- Subtle clues can be used to classify people beyond what human observation can do. Famous example: Google determining that a teenage girl was pregnant before her parents knew.
- People can be more easily tracked by combining clues as to where they are. This aids foreign intelligence agencies facilitation assassinations.

Risk: AI supports totalitarianism

- If it’s easier to track citizens with AI and to tell what they are doing, it is tempting for bad governments to exploit this to control people.
- China and Russia are totalitarian and they like AI. They use it to stifle dissent.
- “Corporate fascism” is increasingly apparent in big monopolies like Google, Microsoft, Facebook, and Amazon to force you to buy their products and hide their flaws. AI supports many of their practices.