Risks of AI systems

Prof. Neil C. Rowe

Computer Science Department
Naval Postgraduate School

http://faculty.nps.edu/ncrowe

Summer 2021

Risk: AI design can be oversimplified

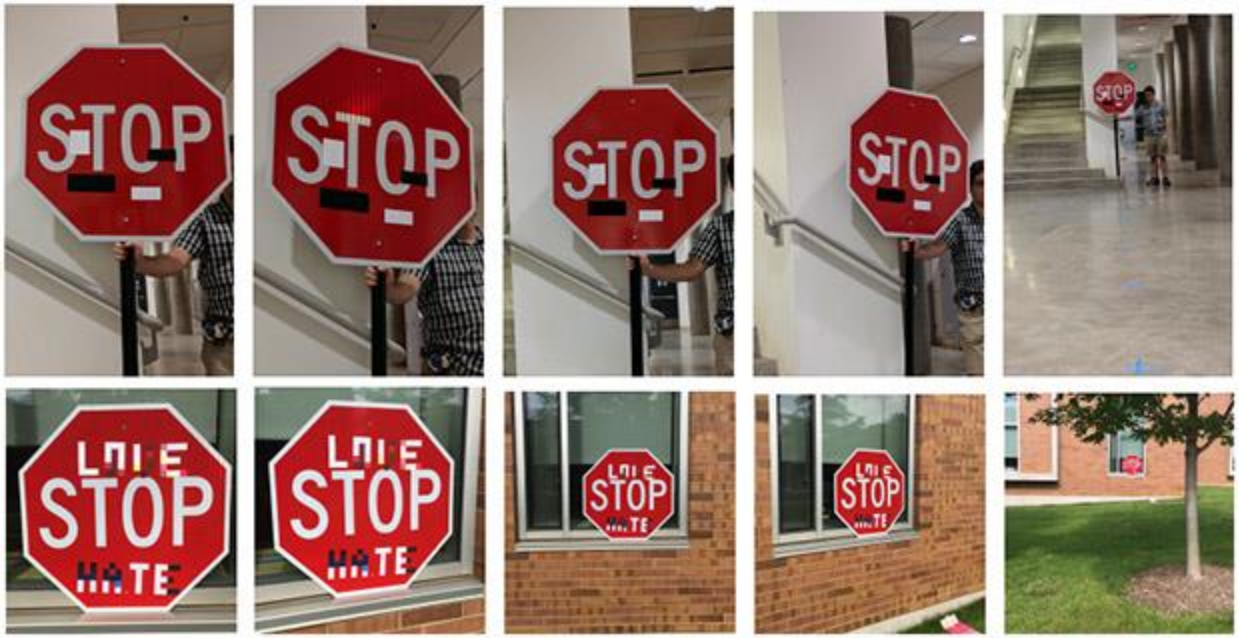
- Driverless cars cannot plan for every rare road situation. So they may cause accidents.
- Similarly, automated machine guns (in Korea) and lethal drones (in the Middle East) may not understand everything they see, e.g. attempts to surrender.
- Financial pressures may encourage vendors to oversimplify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

From: https://en.wikipedia.org/wiki/Self-driving_car

From: From: https://blog.seagate.com/human/only-data-at-the-edge-will-make-driverless-cars-safe/

AI may identify the wrong features

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 image changes that would make it fail. The neural net is clearly not using color and shape as humans do to identify stop signs.



From:
https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms

Risk: AI may be biased

- Bias may occur in a training set for supervised learning, e.g. a skin-disease diagnosis system that was trained on white patients and isn’t as good on black patients.
- Bias may also occur when judgement is used to label data for a training set, as when medical radiology experts prefer diagnoses which require radiology.

Risk: Bugs and errors 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 bugs. Example: a robot car misinterprets a stop sign.
- Legal responsibility for faulty AI should reside with the writers or trainers of the software. But proving responsibility can be difficult.
- Failure of AI to explain itself exacerbates the problem.
- Malicious actors can insert bugs in AI and falsify data.

Risk: Blame becomes harder with AI

- It is harder to see who or what to blame when things go wrong with AI systems.
- Also, 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.

From: https://en.wikipedia.org/wiki/Unmanned_aerial_vehicle

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 false patterns they see.

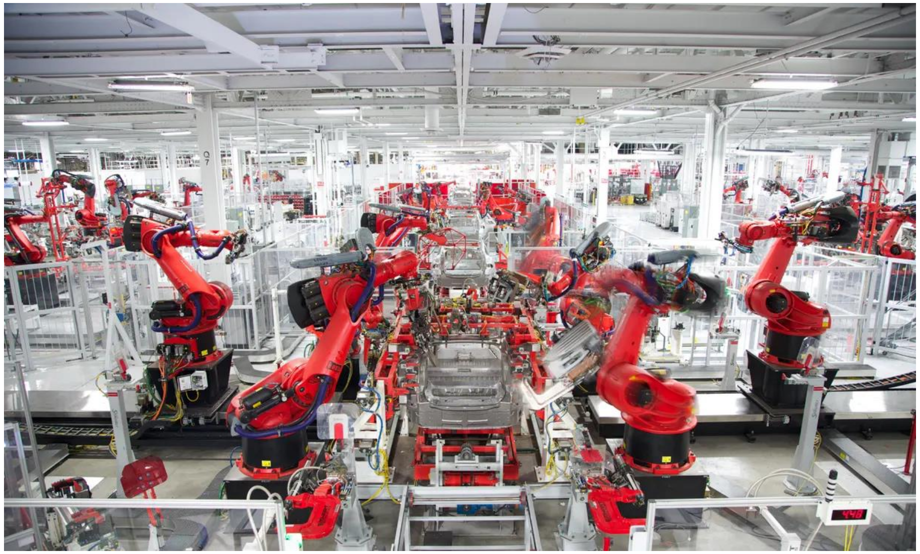
From: https://www.azcentral.com/story/travel/arizona/2017/10/02/ghost-hunting-shows-fake/705566001/

Risk: AI won't know all human needs

- AI creates obedient servants, so it should be considering human needs as primary.
- However, it may hard to identify all the human needs that AI should consider.
- For instance, in an emergency, an AI needs to assign different priorities than normal, and not just continue routine activity.
- Similarly, goal-driven AI may use criteria we don't like, like endangering a few to save many.
- AI may need to lie and cheat to interact effectively with people, like flattering them – is this a good policy in the long run?

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 encourage a society where most people live on government support.



From: https://www.slashgear.com/tesla-model-s-factory-tour-shows-elon-musks-robot-army-17290737/

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 done by AI with business rather than moral considerations. A European law tries to prohibit this.

From:
https://clarionindia.net
/starving-children/

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 determined 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 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 classify activity and stifle dissent.
- “Corporate fascism” is increasingly apparent in big monopolies like Google, Microsoft, Facebook, and Amazon to lock you into their products and ignore their flaws. AI supports many of their practices.

From: https://whyy.org/articles/salute-leader-rules-regs-maga-military-parade/