Responsible Machine Learning: A Primer

Patrick Hall

H_2O.ai

February 13, 2020

Contents

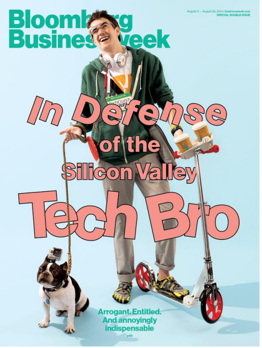
Time to Grow Up

Machine learning (ML) is about 60 years old.

Image: University of Alberta

Then.

Image: Bloomberg


Bloomberg Businessweek
In Defense of the Silicon Valley Tech Bro
Arrogant. Entitled. And annoyingly indispensable

Now.

(So why don't we act like it?)

Risky Business?

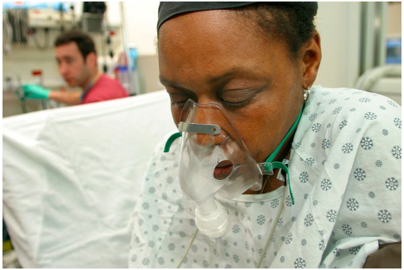
All technologies present risks. ML is no different.

Image: Wikipedia

About 60 years into aviation technology, the Boeing 707 was born. This model of plane has been involved in 261 accidents, causing a total of 3,039 fatalities, over it's long deployment. It was still flying *and crashing* in 2019.

But What Could Go Wrong??!

- Being wrong.
- Discrimination.
- “Computer Says No.”
- Privacy harms.
- Security vulnerabilities.
- Reputational damage.
- Noncompliance.
- Fines and financial losses.
- Systemic failures.

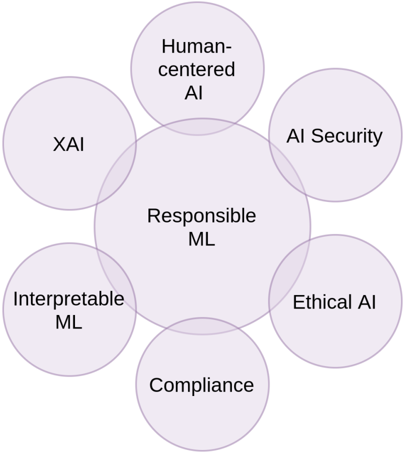


“Millions of black people affected by racial bias in health-care algorithms”
— Nature, 24 Oct. 2019

Vocabulary Quiz

- **Explainable AI** (XAI): the ability to analyze a ML model after it has been developed.
- **Interpretable ML**: transparent model architectures and increasing how intuitive and understandable ML models can be.
- **Ethical AI**: sociological fairness in ML model predictions (i.e., whether one category of person is being weighted unequally).
- **AI Security**: debugging and deploying ML models with similar counter-measures against insider and cyber threats as would be seen in traditional software.
- **Human-centered AI**: user interactions with AI and ML systems.

Responsible ML



Responsible ML is essentially a combination of existing technologies and best practices that a *responsible* organization would employ to mitigate ML risks, including regulatory compliance processes.

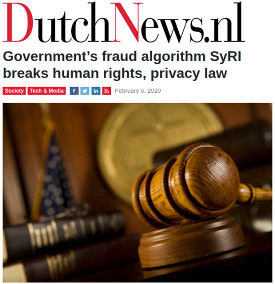
I Fought the Law ...

AI and law are colliding.

Government agencies in *at least* ...

- Canada
- Germany
- Netherlands
- Singapore
- UK
- USA

... have issued or proposed AI-specific guidance.



DutchNews.nl

Government's fraud algorithm SyRI breaks human rights, privacy law

Society Tech & Media February 5, 2020

"Government's fraud algorithm SyRI breaks human rights, privacy law" — dutchnews.nl, Feb. 5 2020

Transparency Paradox

> *"[T]here are two ways of constructing a software design: One way is to make it so simple that there are obviously no deficiencies and the other way is to make it so complicated that there are no obvious deficiencies."*
>
> *— C.A.R. Hoare, 1980 ACM Turing Award Lecture*

- XAI presents privacy risks.
- Interpretable models, XAI, and the model documentation they enable can increase legal liability if they show an organization was aware of a serious problem.
- Consider conducting mission-critical ML projects and risk assessments under legal privilege.

References

Awesome machine learning interpretability metalist
https://github.com/jphall663/awesome-machine-learning-interpretability

An Introduction to Machine Learning Interpretability - 2nd Edition
https://www.h2o.ai/oreilly-mli-booklet-2019

Proposed Guidelines for the Responsible Use of Explainable Machine Learning
https://arxiv.org/pdf/1906.03533.pdf