Responsible AI

Stefan Hermanek

2017: Carnegie Mellon University Masters

2018-19: AppDynamics, PM for Analytics

2019-20: People.ai, PM for ML and Search

2020: Joined Facebook AI Red Team (PM)

2022: Responsible AI Robustness & Safety

Responsible AI Overview

1. Bias
2. Fairness and equity
3. Calibration
4. Provenance
5. Validation and misinformation
6. Robustness and resilience
7. Safety
8. Security
9. Privacy
10. Transparency
11. Interpretability
12. Explainability
13. Empowerment
14. Redress
15. Accountability and governance

FACEBOOK AI Georgia Tech

Define responsible AI concepts (where possible: using examples)

Raise awareness of the issues and responsibilities of an engineer

Objective

AI is a set of technologies that has been **changing the world around us**

We are **building the future with AI**

As practitioners, we **must ensure that we do so responsibly**

Premise

Society, public debate, and governmental **regulation** are on the horizon for AI systems.

- Examples: EU AI Act; regulatory investigations worldwide

As a result, it is not only a **moral imperative** to “do the right thing” and build AI responsibly, but also **good business**. (Hint: job opportunities!)

From an **individual perspective**: Meta (FB) employs thousands of engineers working to combat misinformation and build AI responsibly.

Need and benefits of being responsible

- **Models may have a bias** – we talked about this in bias-variance tradeoffs
- **Bias in common language:** Any form of preference, fair or unfair
- **Bias in the context of responsible AI:** “Unfair”, “unwanted”, or “undesirable” bias[1]

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

	Gender stereotype *she-he* analogies	
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant
	Gender appropriate *she-he* analogies	
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Examples of gender biases in word embeddings[2]

There is a spectrum of fairness and equity – from “basic” measures to more justice-focused outcome-oriented equity

Fairness:

- There exist multiple criteria to define fairness
- Ask: Fairness with respect to “which dimension” (e.g. gender, age, etc.)

Equity:

- Goes beyond equality
- Employs a “justice approach [to] conside[r] how certain groups are oppressed or marginalized in the particular context and explores how the AI system can advance equity, rather than perpetuate a status quo that may oppress or marginalize certain groups.” [3]

- **COMPAS** = commercial system for predicting recidivism (re-offense), widely used in the US legal system[4]

- **"Individuals who are given the same score […] have approximately the same probability of re-offending"**
  - Risk score of 7/10 → "60% of whites and 61% of blacks re-offend"
  - Regardless of race

- This system is well calibrated. **But is it fair?**

- "proportion of those who did not re-offend but were falsely rated as high-risk was 45% for blacks and 23% for whites" **(i.e. false positive rate difference)**

- There is an **inherent trade-off** between an algorithm being well calibrated and equal outcomes for different groups.

- **Challenge for practitioners** – how to draw that tradeoff? What to optimize for?

- **Know thy data (and its usage):**
  - Where did data arise?
  - What inferences were drawn from the data?
  - How relevant are those inferences to the present situation?

- Term borrowed from **database research**

- **Provenance matters** for auditability, explainability, and debuggability of an AI system.

- It matters especially in **highly complex, chained, and distributed AI systems.**

Validation

- **Software:** Ensure your code runs as intended (predictably so)
- **AI:** Ensure your model performs as designed (predictably-ish so)

There are various ways of performing validation:

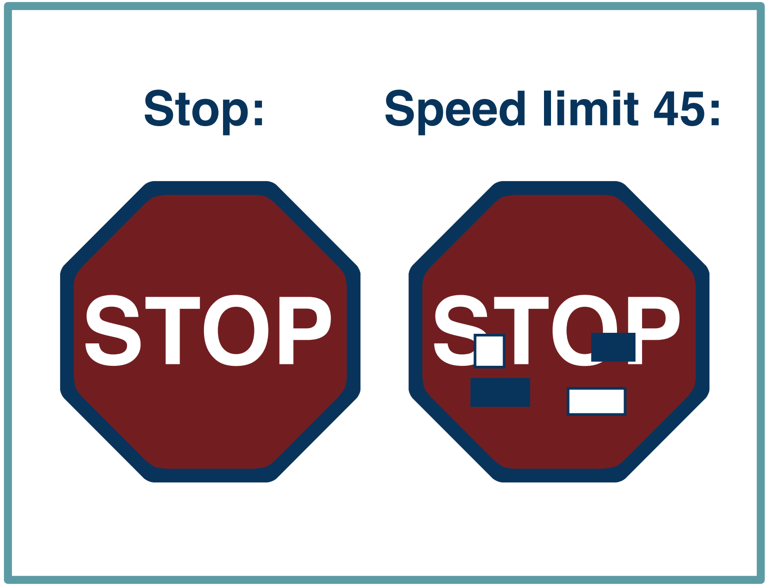
- **In classic software development:**
  - Unit tests, integration tests, regression tests, or perform penetration tests
  - Penetration tests may find zero-days – privately-known vulnerabilities “out in the wild”
- **For AI systems:**
  - Validation tests (train/test/validate), e.g. using k-fold cross-validation.
  - Red Teaming AI systems, using penetration tests and adversarial approaches
- **But what about “claims” – expressions of statement like (“the sky is green”)?**
  - Fact checking claims has been an approach to combat misinformation online

Robustness = "building reliable, secure ML systems" [6]

- Systems that perform in a reliable manner…
- …over time…
- …and as designed…
- …under reasonable (known) conditions

Resilience = "ability to adapt to risk" [7]

- Systems that perform predictably…
- …when faced with novel situations (e.g. in adversarial scenarios)
- …and let owners know when that happens.



Lack of resilience in AI systems[8]

Robustness and Resilience

- **Safety is domain-specific**
  - Predicting {hotdog | no-hotdog} will have other safety requirements than self-driving cars.
- **Audit systems** for potential safety issues and potential for harm
- **Techniques:**[4]
  - Failure modes and effect analysis (FMEA) – what could go wrong and how?
  - Fault tree analysis (FTA) – what conditions lead to failure modes?
- Harm goes beyond direct harm – **unintended side effects** (e.g. polarization) and **externalities** (e.g. environmental impact)

- **AI systems are software systems** – the same cybersecurity paradigms still apply (e.g. encrypting data in transit, authentication, access control)

- **AI systems are unique** – and raise specific security concerns:
  - **Poison training data** (similar to supply chain corruption)
  - **Evade a model** (stop sign example as “adversarial attacks”, model-based attacks)
  - **Exfiltrate data or information from the model** (membership inference, data exfiltration)
  - …

Security

Ensure that subjects' data remains private (both in development and inference)

Approaches:[4]

- **De-identify instances**
  - Example: remove names
  - Challenge: re-identification often easy, e.g. by age, zip-code, social security number, etc.
- **Generalize fields**
  - Example: generalize information to make individual records anonymous, e.g. ">60 years"); continue until at least k records are indistinguishable (k-anonymity)
  - Challenge: individual records still exist; requires trust
- **Query in aggregate**
  - Example: return only results that satisfy certain conditions, e.g. k-anonymity
  - Challenge: individual records still exist; requires trust

Ensure that subjects' data remains private (both in development and inference)

Approaches:[4]

- **Differential privacy**
  - Intuition: Add noise to query results so that re-identification is impossible
  - Challenge: information loss; requires trust
- **Federated learning**
  - Intuition: Split data / model between model owner and data subject (e.g. server and end-user device)
  - Challenge: information loss; requires trust; requires compute
- **Secure hardware for privacy-preserving processing of data**
  - Secure processors (a smartphone has one; cloud computing services have many thousands)
  - Hardware security modules (credit-card readers have them, as do banks and ATMs)

Transparency is deeply intertwined with explainability

- Why did a model make a prediction?
- How did it arrive at its decision?

Transparency goes beyond explainability:

- What data was used to train a model?
- What is a model's intent?
- When was the model last trained?

Transparency poses unique challenges:

- Competitive considerations
- Robustness/safety tradeoff – "the more an adversary knows, the more they can use that knowledge"

Larger models + richer feature representations → How does a particular system work?

Interpretability generally asks questions at the "model/system"-level:

- Which features are important?
- Which are not important?

Larger models + richer feature representations → How did a particular prediction come about? What went into it?

Interpretability generally asks questions at the "instance"-level:

- Why was this particular prediction made as it was made?
- Which features led to a particular decision?

Dictionary definition:[9]

- "the act or action of empowering someone or something : the granting of the power, right, or authority to perform various acts or duties"
- "the state of being empowered to do something : the power, right, or authority to do something"

What does this mean for AI systems?

- Educate users about usage of AI systems
- Provide users with controls in interacting with AI systems (e.g. manual over-ride by a judge)

Dictionary definition:[10]

- “relief from distress”
- “means or possibility of seeking a remedy”
- “compensation for wrong or loss”

What does this mean for AI systems?

- Premise: As “trained” statistical models, AI systems are inherently going to get it wrong.
- It’s imperative to provide users with means of rectifying harm.

Redress

Accountability = “the state of being responsible or answerable for a system, its behavior and its potential impacts”

Challenge: algorithms are not moral or legal entities. Then, who is accountable?

- The organizations and people building and deploying AI systems
- How? Governance

Governance = a process for ensuring accountability, compliance, and ethical decision making when building AI systems

References

1. Silberg, J., & Manyika, J. (2019, June). *Notes from the AI frontier: Tackling bias in AI (and in humans)*. Tackling Bias in AI. Retrieved November 4, 2021, from https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf.
2. Bolukbasi, T., Kai-Wei, C., James, Z. Y., Saligrama, V., & Adam, K. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems 29 (NIPS 2016)*. Retrieved November 4, 2021, from https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
3. Smith, G. (2020). *What does "fairness" mean for Machine Learning Systems?* What is Fairness? Retrieved November 4, 2021, from https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf.
4. Russell, S. J., Norvig, P., & Chang, M.-W. (2022). Chapter 28: Philosophy, Ethics, and Safety of AI. In *Artificial Intelligence: A modern approach* (pp. 1032–1062). essay, Pearson.
5. Jordan, M. I. (2019). Artificial Intelligence—The Revolution Hasn't Happened Yet. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.f06c6e61

References

6. Steinhardt, J., & Toner, H. (2020, August 24). *Why robustness is key to deploying AI*. Brookings. Retrieved November 4, 2021, from https://www.brookings.edu/techstream/why-robustness-is-key-to-deploying-ai/.
7. Petrilli, A. , & Lau, S. , (2019, December 12). Measuring Resilience in Artificial Intelligence and Machine Learning Systems [Blog post]. Retrieved from http://insights.sei.cmu.edu/blog/measuring-resilience-in-artificial-intelligence-and-machine-learning-systems/
8. Heaven, D. (2019). Why deep-learning AIS are so easy to fool. *Nature*, *574*(7777), 163–166. https://doi.org/10.1038/d41586-019-03013-5
9. Merriam-Webster. (n.d.). Empowerment. In *Merriam-Webster.com dictionary*. Retrieved November 4, 2021, from https://www.merriam-webster.com/dictionary/empowerment
10. Merriam-Webster. (n.d.). Redress. In *Merriam-Webster.com dictionary*. Retrieved November 4, 2021, from https://www.merriam-webster.com/dictionary/redress
11. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372873

References