Risks with AI

- Overfitting
- Social and ethical risks

Risks With AI: Overfitting

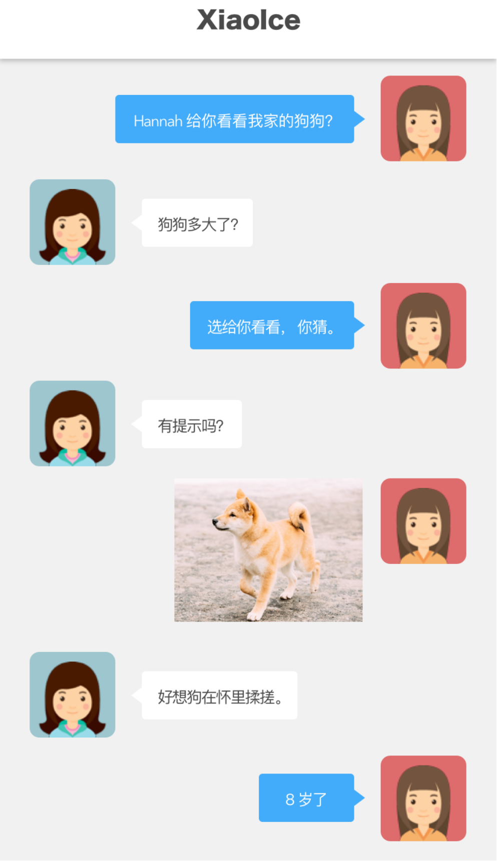
- Complex AI models such as Neural Nets can easily overfit (i.e. fit historical data too well but fail in realistic test conditions). If we don't understand what is helping the model perform well, there is a risk that the model will fail upon deployment
- Operational Risks
  - Direct financial risks (e.g. a trading algorithm)
  - Customer perception & reputation (e.g. poor personalization experience)
- ML Models need to go through thorough stress-testing (discussed later under audits)

Xiaoice: Darling of Chinese Social Media

XIAOBING

40M followers

China



XiaoIce

Hannah 给你看看我家的狗狗?

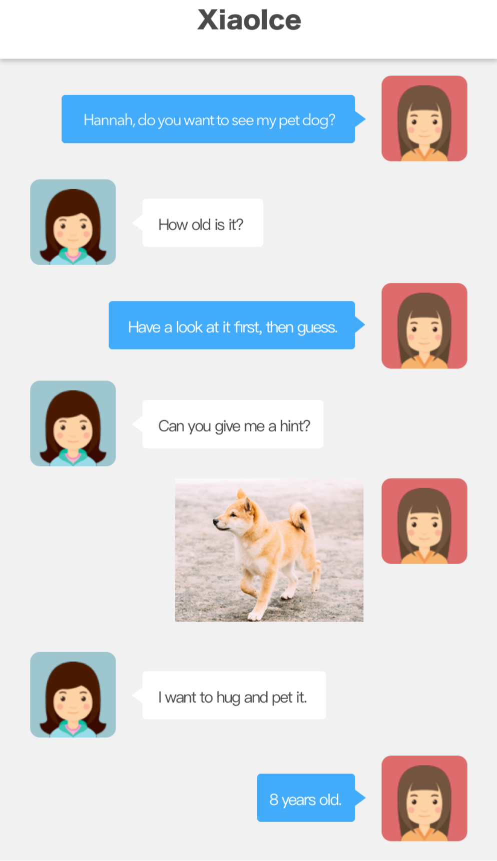
狗狗多大了?

选给你看看，你猜。

有提示吗?

好想狗在怀里揉搓。

8 岁了



XiaoIce

Hannah, do you want to see my pet dog?

How old is it?

Have a look at it first, then guess.

Can you give me a hint?

I want to hug and pet it.

8 years old.

YUAN ZHANG

22 years old

China

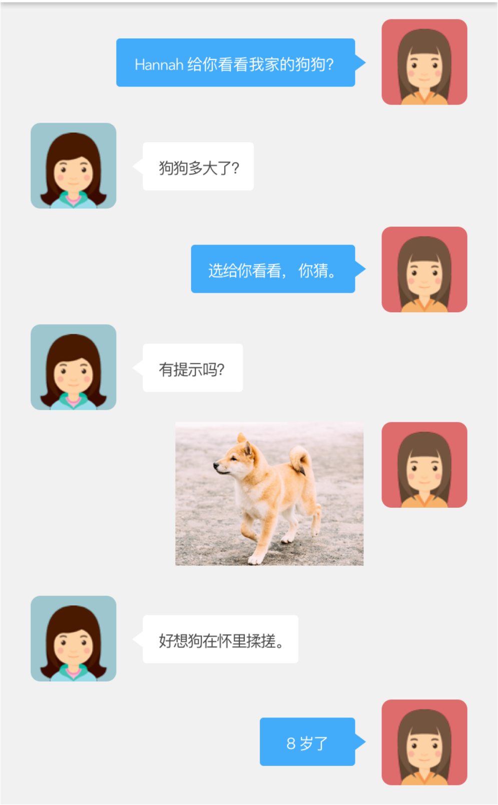
Xiaoice: Darling of Chinese Social Media

XIAOICE

40M followers

Chatbot

XiaoIce



Hannah 给你看看我家的狗狗?

狗狗多大了?

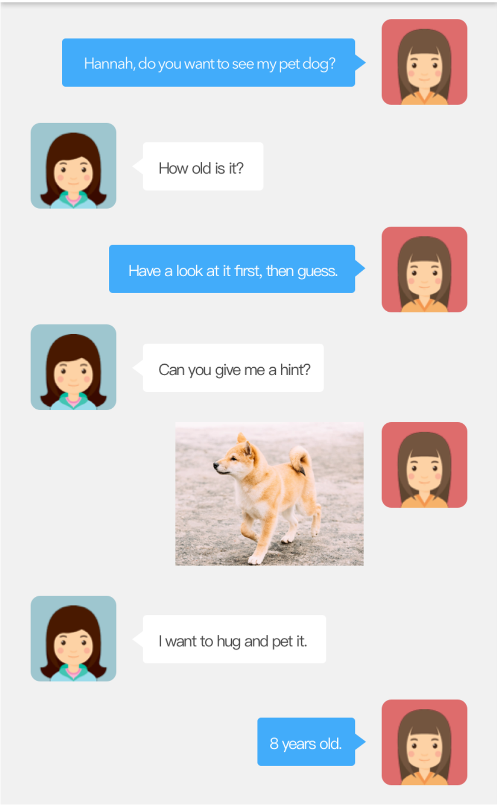
选给你看看，你猜。

有提示吗?

好想狗在怀里揉搓。

8 岁了

XiaoIce



Hannah, do you want to see my pet dog?

How old is it?

Have a look at it first, then guess.

Can you give me a hint?

I want to hug and pet it.

8 years old.

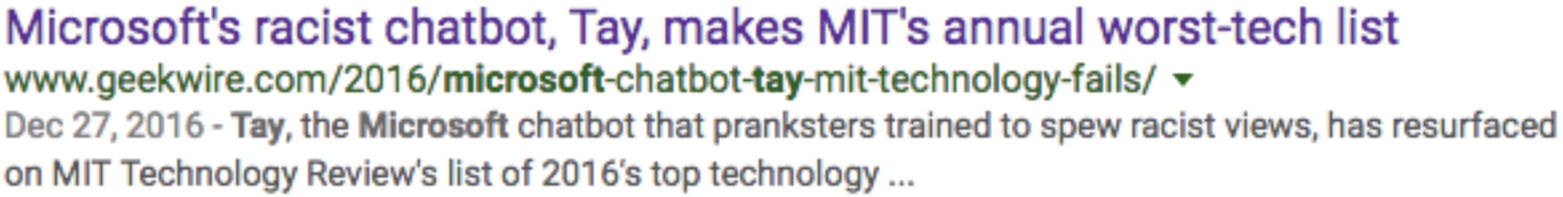
YUAN ZHANG

22 years old

China

Wharton
University of Pennsylvania

Tay.ai: Xiaoice's Evil Cousin


Microsoft's racist chatbot, Tay, makes MIT's annual worst-tech list

www.geekwire.com/2016/**microsoft**-chatbot-**tay**-mit-technology-fails/ ▾

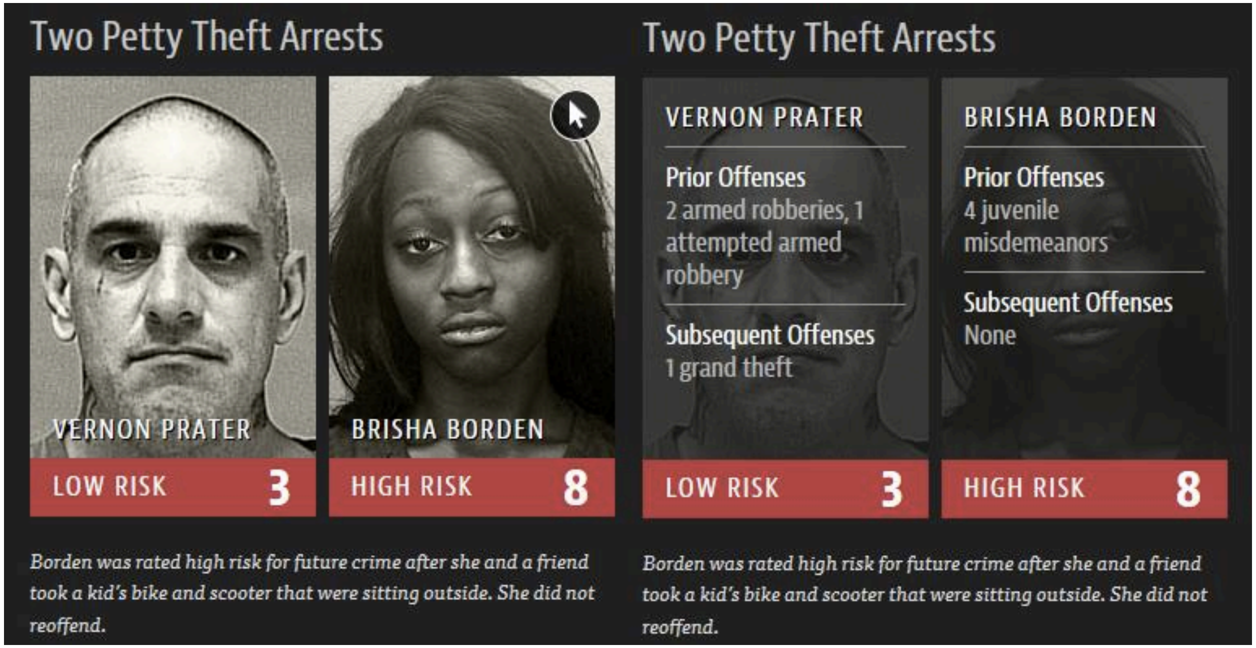
Dec 27, 2016 - **Tay**, the **Microsoft** chatbot that pranksters trained to spew racist views, has resurfaced on MIT Technology Review's list of 2016's top technology ...

Algorithm Bias in Recruiting Software

Amazon’s machine learning specialists uncovered a big problem: their new recruiting engine **did not like women**.

Algorithms Incorrectly Predict Recidivism

There are **higher false positive and false negative rates** among African Americans.



Two Petty Theft Arrests

VERNON PRATER

LOW RISK 3

BRISHA BORDEN

HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Algorithms racial bias in predicting recidivism rates.

Why Might AI-based Decisions Unpredictable?

Algorithm Logic

Nature

Data

Nurture

AI Behavior

Analogous to human behavior, algorithms are driven by **nature and nurture.**

Risks to Society

- Social risks can result from automated decisions because these decisions may result in disadvantaged minorities continuing to be left behind.
  - Biases exist in the world and input data often contains these biases, so algorithms can incorporate discrimination and exacerbate it.
  - The harms that result can be categorized into the following two groups, originally described by Kate Crawford of the AI Now Institute.

Harms of Allocation	Harms of Representation
• "When a system allocates or withholds opportunities or resources to or from certain groups" • This affects what people can access • E.g.: Unfairly denying access to a loan, mortgage, or job on the basis of gender, race, age, etc.	• "When a system misrepresents a group in a damaging way" • This affects how people are viewed • E.g.: Black faces labeled as gorillas by Google Photos; TSA scanners being more likely to false alarm people of color (based on hairstyles).

Content/quotes from: https://medium.com/@jstanier/we-must-fix-ais-diversity-problem-6ad5fddc2f8c
Content from: https://gizmodo.com/microsoft-researcher-details-real-world-dangers-of-algo-1821129334

Risks to Firms

- These social risks then create additional risks for companies:

REPUTATIONAL RISK	LEGAL RISK	REGULATORY RISK
• Perceived to be a biased, prejudiced company • Firms may face PR issues and backlash as a result	• Sued for unfair practices and discrimination against particular groups	• Increased regulation & cost of compliance • Upcoming interest in auditing and data protection (GDPR)

AI Governance

- User Control
- Transparency
- Audits

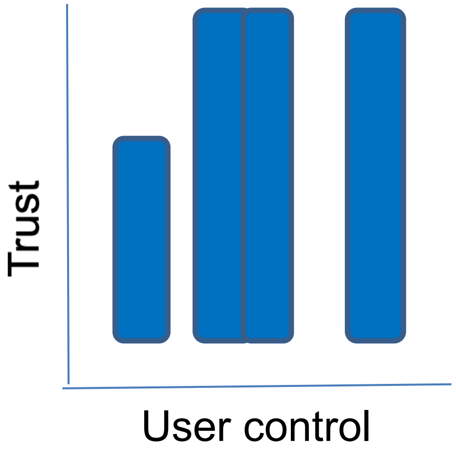
User Control

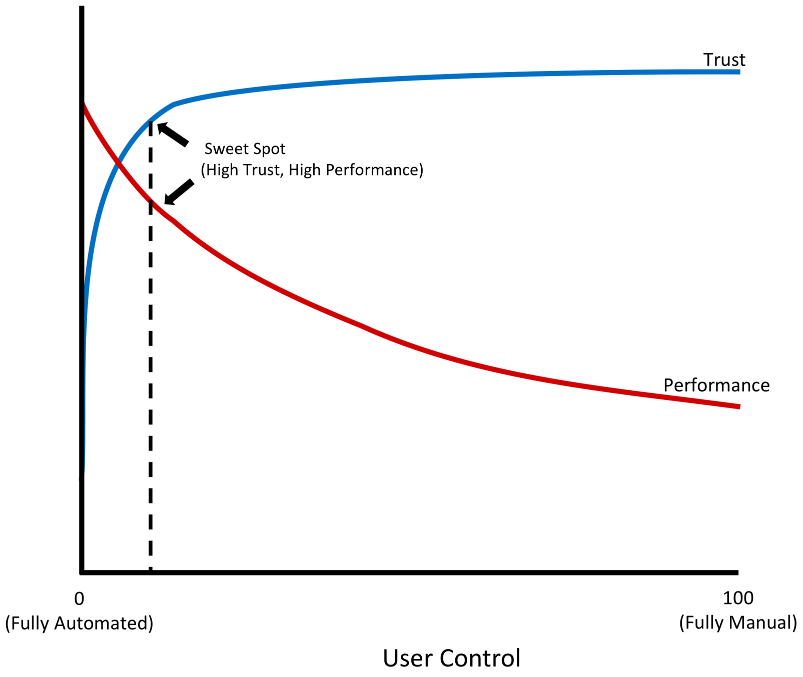
- Giving users some control over the decisions algorithms make for or about them
  - E.g., Facebook newsfeed now allows users to flag posts in their newsfeed as being false or offensive
  - System has helped detect many problematic posts
- In 2015, Facebook released mixed-style newsfeed controls
  - Allowed users to decide whether they wanted more/less of particular news (relationship statuses, profile changes, specific friends, etc.)
  - User satisfaction did increase (people liked having control)
  - But engagement went down (the algorithm knew what would engage users better than the users themselves)
- Demonstrates that there is a role for providing user control, but it needs to be balanced with ensuring the algorithm performs well.

Content/quotes from: “A Human’s Guide to Machine Intelligence” by Kartik Hosanagar

User Control

- There is a sweet spot that companies should aim for: providing a low level of control that enhances user trust, but not so much control that it significantly affects algorithm performance.





Content/quotes from: “A Human’s Guide to Machine Intelligence” by Kartik Hosanagar

Transparency for End Users

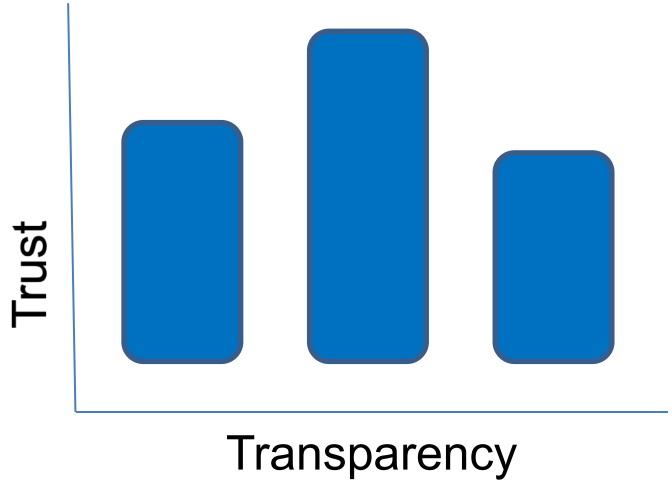
- Users can also be uncomfortable with decisions made by algorithms if they feel they don't understand those decisions.
- Solutions to this sometimes focus on **technical transparency**
  - Technical transparency = revealing source code
  - Examples: 2015 CFTC ruling on 2010 flash crash, 2017 NYC proposed bill on automated decisions
  - Problems with technical transparency: Doesn't protect intellectual property, creates vulnerability to adversarial attacks, doesn't always explain decisions (in cases of deep learning, random forests, etc.)

- **Moreover, technical transparency may not even be necessary or helpful in improving user trust.**

Content/quotes from: "A Human's Guide to Machine Intelligence" by Kartik Hosanagar

Transparency for End Users

In a study on grading by Rene Kizilcec, students who were provided a limited amount of transparency had the greatest trust in their grades. Those who were given more extensive explanations of how the algorithm functioned had trust levels similar to those with no explanation at all.



"Calibrated transparency" is a better approach:

- Was an algorithm used to make a decision?
- What kinds of data are used?
- What variables are considered?
- Global and local interpretability

Content/quotes from: "A Human's Guide to Machine Intelligence" by Kartik Hosanagar

Transparency for Managers and Data Scientists

- Many of the best performing ML models are often highly opaque (e.g. Neural Network making loan approval decisions)
- Recent interest in Interpretable ML (also known as Explainable AI)
  - Global interpretability: Can we explain at a high level what are the most important variables driving a model's predictions (e.g. income, credit history, etc)
  - Local interpretability: Can we explain the most important variables driving a particular prediction or decision (why was Kartik's loan application not approved?)
- Many open source tools and third-party vendors are starting to offer model interpretability solutions
  - Can be valuable for debugging

Auditing Algorithms

- Due to the risks automated decisions bring, ML should be designated as a distinct model type, with it's own governance frameworks.
- The Algorithmic Accountability Act could be one such framework
  - If passed, it would require large companies to formally evaluate their "high-risk automated decision systems" for accuracy and fairness. But forward-thinking companies should not wait for regulation.
- What would an audit look like?
  - An audit process would begin with the creation of an inventory of all machine learning models being employed at a company as well as:
    - The specific uses of these models
    - The names of the developers & business owners of models
    - Risk ratings: the social/financial risks if the model fails

Wharton
UNIVERSITY of PENNSYLVANIA

Auditing Algorithms

- For high-risk models, the audit should look at the following three areas:

Inputs	Model	Outputs
• Data quality • Bias in training data	• Alternative models • Statistical tests for model fit, overfitting • Model transparency • Stress test against simulated data	• Decisions with explanations • Outliers: range of inputs and outputs

- Machine learning models involve a variety of complex issues, such as bias, interpretability, and the fact that they are constantly retraining as they receive more data.

Three Lines of Defense

Three Lines of Defense for Data Science

Model Developer

Data Science QA

Data science auditor
(for high-stakes models)

Wharton

UNIVERSITY of PENNSYLVANIA