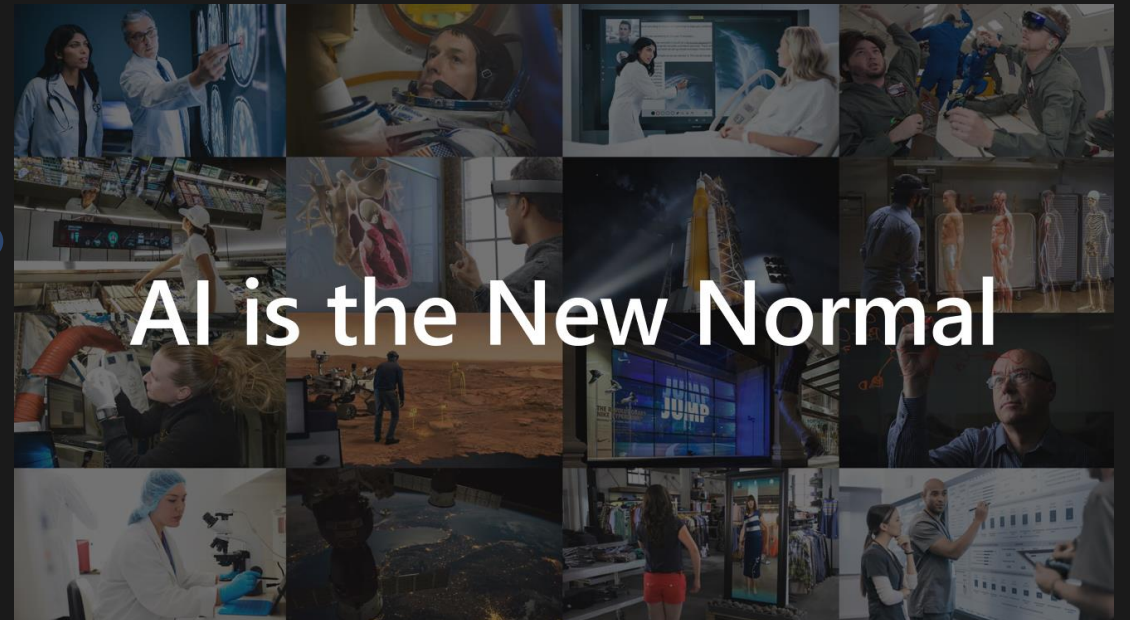




Vertrauenswürdige KI-Dienste

→ *Responsible AI*

Anja Fiegler, AI Solution Architect
Strategy & Innovation, Microsoft Germany



THE WALL STREET JOURNAL.

English Edition | June 12, 2020 | Print Edition | Video | Latest Headlines

Home World **U.S.** Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ Magazine

WSJ | CIO Network *How can I... keep our customers safe and our network secure?* [Learn More](#)

U.S. | NEW YORK

New York Regulator Probes UnitedHealth Algorithm for Racial Bias

Financial Services Department is investigating whether algorithm violates state antidiscrimination law

DATAVERSITY

Home Conferences Online Training Live Webinars White Papers Product Demos

Homepage > Data Education > Smart Data News, Articles, & Education > The Illusion of Data Anonymization

The Illusion of Data Anonymization

By A.R. Guess on May 19, 2011

[Twitter](#) [Facebook](#) [LinkedIn](#)

Pete Warden recently commented on anonymized data, stating, "One of the joys of the last few years has been the flood of real-world datasets being released by all sorts of organizations. These usually involve some record of individuals' activities, so to assuage privacy fears, the distributors will claim that any personally-identifying information (PII) has been stripped. The idea is that this makes it impossible to match any record with the person it's recording. Something that my friend Arvind Narayanan has taught me, both with theoretical papers and repeated practical demonstrations, is that this anonymization process is an illusion."

Zeitung

Magazin

Themen: Kultur Gesellschaft Wissen Reise Auto mehr...

50 - Die Zukunft

Boter: Mit boardiert

FASTCOMPANY

03-02-19

Here are the data brokers quietly buying and selling your personal information

You've probably never heard of many of the data firms registered under a new law, but they've heard a lot about you. A list, and tips for opting out.

[Source image: ksenia_bravo/istock]

MEINE NEWS | HOME POLITIK UNTERNEHMEN TECHNOLOGIE FINANZEN

Börsenkurse Märkte Anlagestrategie Banken + Versicherungen Geldpolitik

Handelsblatt > Finanzen > Banken + Versicherungen > Apple Card: Weniger Kredit für Frauen?

HEALTH IT SECURITY

Intelligence Healthcare Media

Home News Features Interviews Podcasts Research White Papers & Webinars

HIPAA and Compliance Cybersecurity Cloud Mobile Patient Privacy Data Breaches Privacy

UPDATE: The 10 Biggest Healthcare Data Breaches of 2020, So Far

Despite the COVID-19 crisis, phishing campaigns, mishandled health record disposals, and sophisticated cyberattacks are behind some of the biggest healthcare data breaches of 2020.

heise online

IT Mobiles Entertainment Wissen Netzpolitik Wirtschaft Journal

TOPTHEMEN: CORONAVIRUS E3 HOME OFFICE SMART GARDEN E-AUTO WINDOWS 10

heise online > News > 10/2018 > Amazon: KI zur Bewerbungsprüfung benachteiligte Frauen

Amazon: KI zur Bewerbungsprüfung benachteiligte Frauen

Eigentlich wollte Amazon eine Software entwickeln, die unter Bewerbern automatisch die besten findet. Der Algorithmus hatte aber unerwünschte Nebenwirkungen.

Lesezeit: 1 Min. In Pocket speichern

(Bild: metamorworks/Shutterstock.com)

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

DISKRIMINIERUNG

Kritik an Apple Card: Weniger Kredit für Frauen?

Apple gerät auf Twitter scharf in die Kritik: Nicht nur die Ehefrau des Apple-Mitgründers Steve Wozniak erhielt bei der Apple Card weniger Kredit.

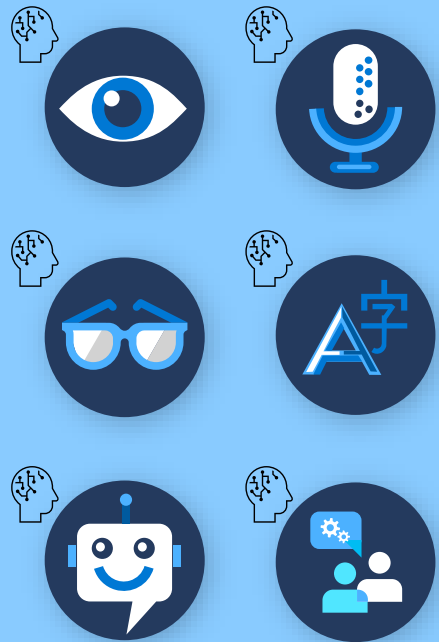
Agenda for Today

- 1 Introduction
- 2 Responsible AI Analysis
- 3 Responsible AI Practices
- 4 Responsible AI Tools
- 5 Discussion & Food for Thought

Artificial Intelligence

Artificial Narrow Intelligence (ANI)

we are here



Predict, create and act

Supervised Learning,
Unsupervised Learning,
Reinforcement Learning

ML / Deep Neural Nets

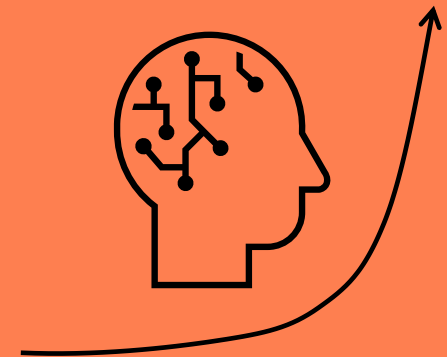
Machine expertise
at a specific task

Artificial General Intelligence (AGI)



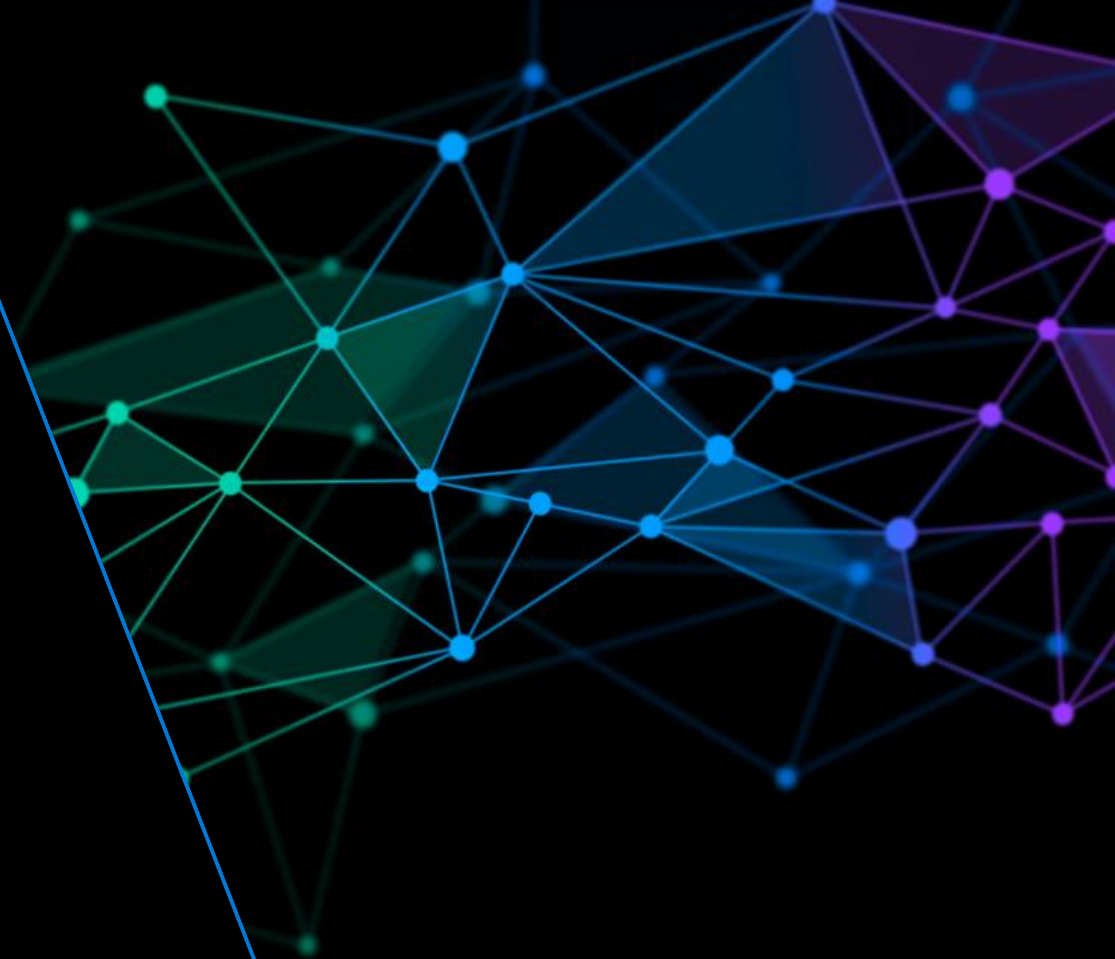
A single algorithm that performs
well on many tasks

Artificial Super Intelligence (ASI)



AI markedly outperforms
human intellectual capabilities

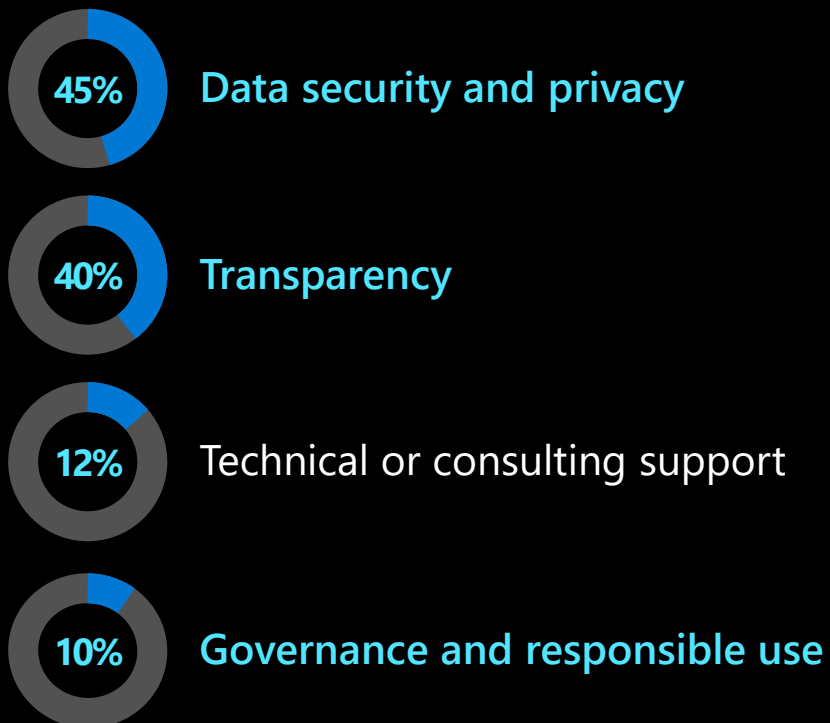
Responsible AI Analysis



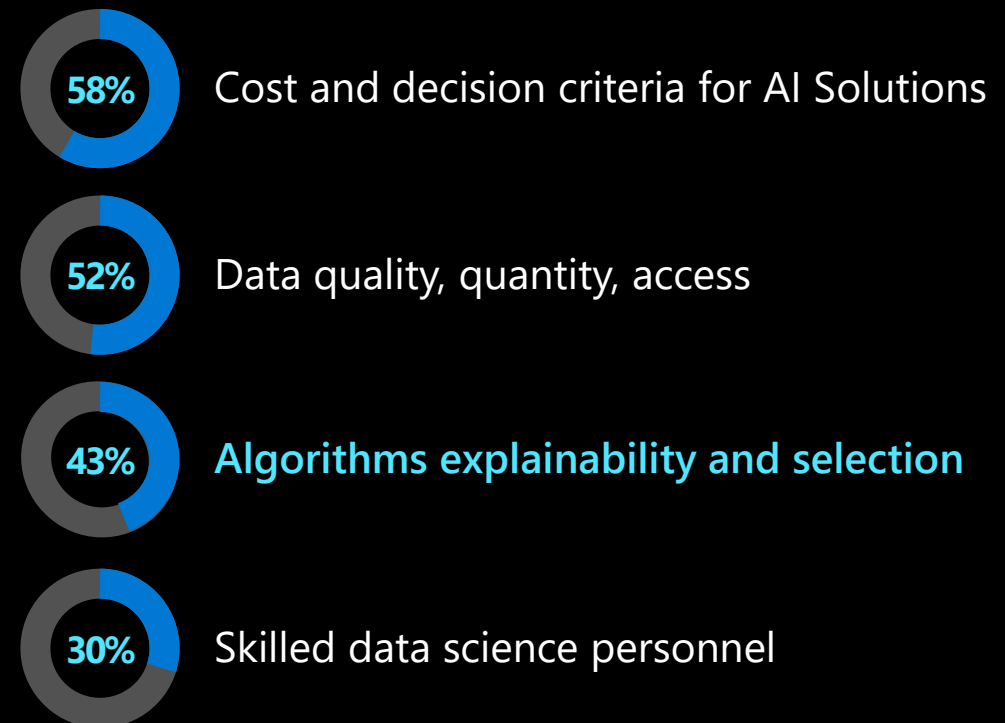
Responsible innovation is top of mind

Most important considerations when investing in AI and Machine Learning technology

Requirements in investing in Machine Learning



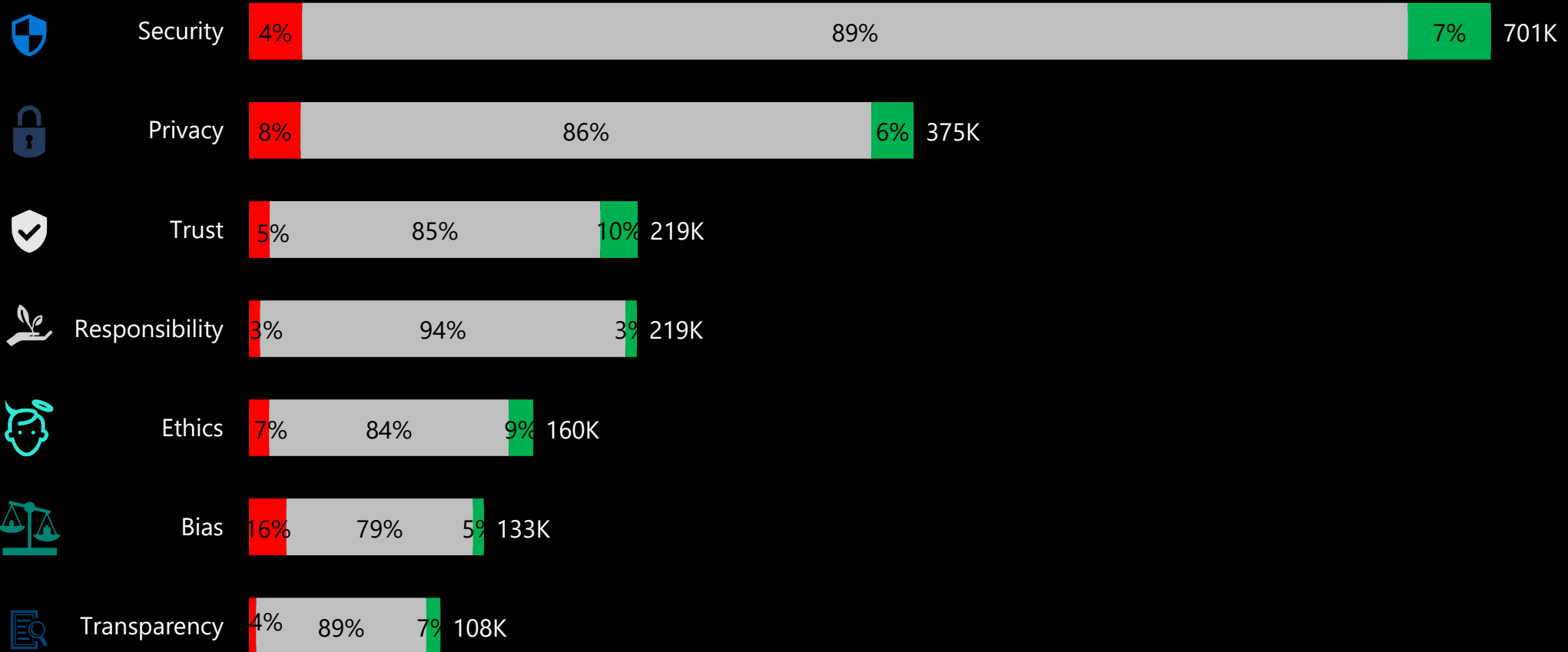
Factors holding businesses from implementing AI



Responsible AI Market Research (Twitter)

Security and privacy drive the overall AI Trust discussion

Volume by Subtopic



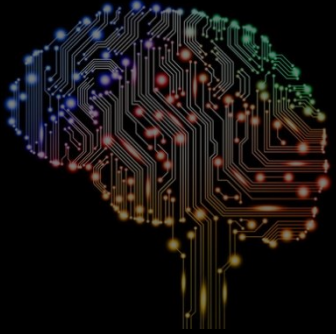
Negative

Neutral

Positive

Time Period: Jul 2018 – Oct 2018

Users distrust AI when they hear of bias issues and don't understand how AI works



Users question the objectivity of AI amid high-profile examples of AI bias and dislike the black-box nature of the AI algorithms – helping users understand how AI works may help build trust in AI.



Bias

Conversation around algorithmic bias is often politically charged, even more strongly impacting trust.



Technology and **AI are not neutral!** There is **bias in every data set** and every algorithm. Ethical considerations and Trust become value creators, facilitated by technology.



Facebook rolling back controversial initiative to fight fake news via the @FoxNews App SURE FACEBOOK. HEADS UP FACEBOOK SPREADS FAKE NEWS WITH THEIR **LIBERAL BIAS** & FAVORABLE **LIBERAL ALGORITHMS**. DO NOT TRUST FACEBOOK, TWITTER OR ANYONE IN THE MSM...LIARS



73% of users believe search engine results are trustworthy, but **gender and racial bias** are embedded in algorithms. @ProfCatherine hosted @SafiyaNoble for a conversation on what that means for digital human rights.



#AI results are only as good, honest & accurate as the humans who build the algorithms. **Can bias & discrimination be built in?** Of course! What independent expert will **validate algorithms** are accurate & fair? This must be part of any AI solution to have trust in it.



Transparency

The complexity of AI and the black box nature of certain machine learning techniques such as neural networks unnerve both users and developers.



The biggest barrier to user trust for AI based systems is **explainability**. **Black box** models don't make sense so each time they're wrong, trust is eroded. **A well explained model** makes mistakes that make sense. #DataScience #MachineLearning #DeepLearning #artificialintelligence



Because few people are going to trust a **black box** to make decisions for us. #artificialintelligence



Most finance leaders **don't understand how #AI works**, so they find it difficult trust to its recommendations. #GRC and auditing can help build responsible AI. @PwCAdvisory outlines how



The future of #ArtificialIntelligence depends on #Trust – if it is to drive business success, #AI cannot hide in a **black box**



IBM announced new **trust and transparency capabilities for AI** to help your business achieve visibility into #AI and deliver more fair, accurate outcomes.

Users distrust AI when performance falls short of expectations and when they don't feel like they have control of it



Users want AI to be error-free by human standards and use highly visible applications such as AutoCorrect as a gauge of AI's overall accuracy.



Accuracy

Users share cautionary tales about placing too much trust in flawed AI.



*I **don't trust any music algorithm** unless it first proves itself by voting Paul Anka's 'She's Having My Baby' as the worst song ever.*



*My most routine daily encounter with AI is **autocorrect**. It **doesn't fill me with confidence**.*



*3rd party certification of AI to gain customer trust? Interesting article - not sure I fully agree but definitely agree with statements like "Whether the use of cognitive technologies is internal or external, it's **best to under-promise and over-deliver**"*



*Oh the irony of someone from IBM's AI unit talking about trust. The first step to acquiring this trust is **delivering what you say you will deliver**.*



*@IBM To answer your question... No. We will **never be able to trust AI** until we can get the autocorrect thing figured out.*



Control

Distrust in people and organizations who develop or use AI often extends the AI itself



*@user I am **scared** to make that leap cause **unless I programmed** that bot personally (which I don't know how to do) I have no way to trust its working for me and not the **bot creator**. But I totally want to if I could find a way to trust it.*



*@guardian Whereupon **Murdoch** hires The **Russian Bot Army** to game **Facebook algorithm** for granting "**Trusted Source**" status in favor of Fox and its subsidiary mouthpieces (Hannity, Cavuto, etc.) and, voila, cash flow. Bot Army will probably do the dirty on commission.*



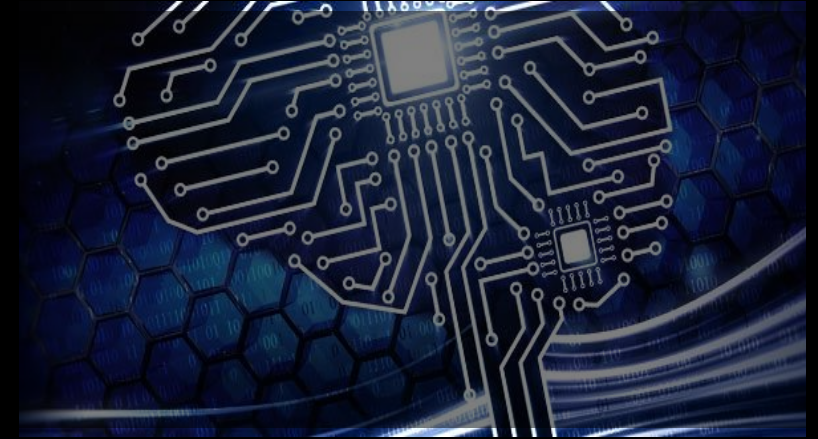
*I pulled ALL my money out of Morgan! **I don't trust robots** to make my **financial decisions**!*



*FACEBOOK USES **ARTIFICIAL INTELLIGENCE TO PREDICT** YOUR FUTURE ACTIONS FOR ADVERTISERS, SAYS CONFIDENTIAL DOCUMENT **#DeleteFacebook***

Both users and industry insiders weigh in on trust in AI

Users question whether AI and its providers can be trusted, while industry insiders discuss how to build trust in AI



KEY THEMES



AI Performance

The AI product experience and whether it delivers on user expectations



The Human Element

The human-like AI product interface and the human creators/owners of AI



AI-Powered Trust

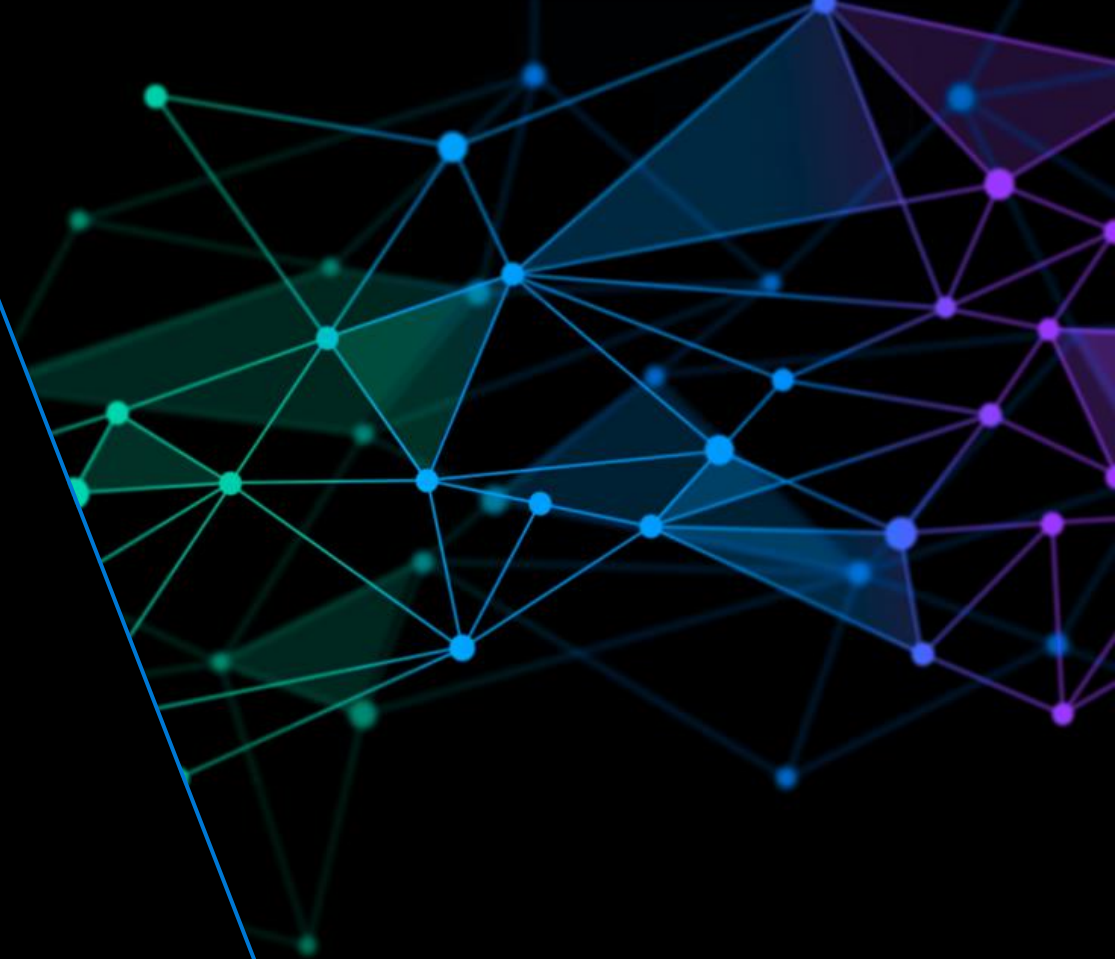
Using AI to evaluate trustworthiness and improve trust in other products and services



Industry Awareness

AI industry insiders discuss the importance of and strategies for building trust

Responsible AI Practice



Put **responsible** AI into Action

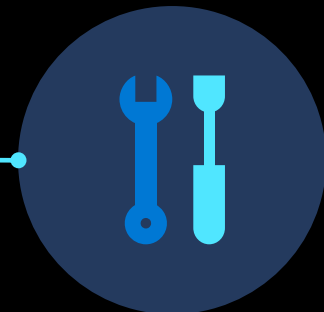
Principles



Practices

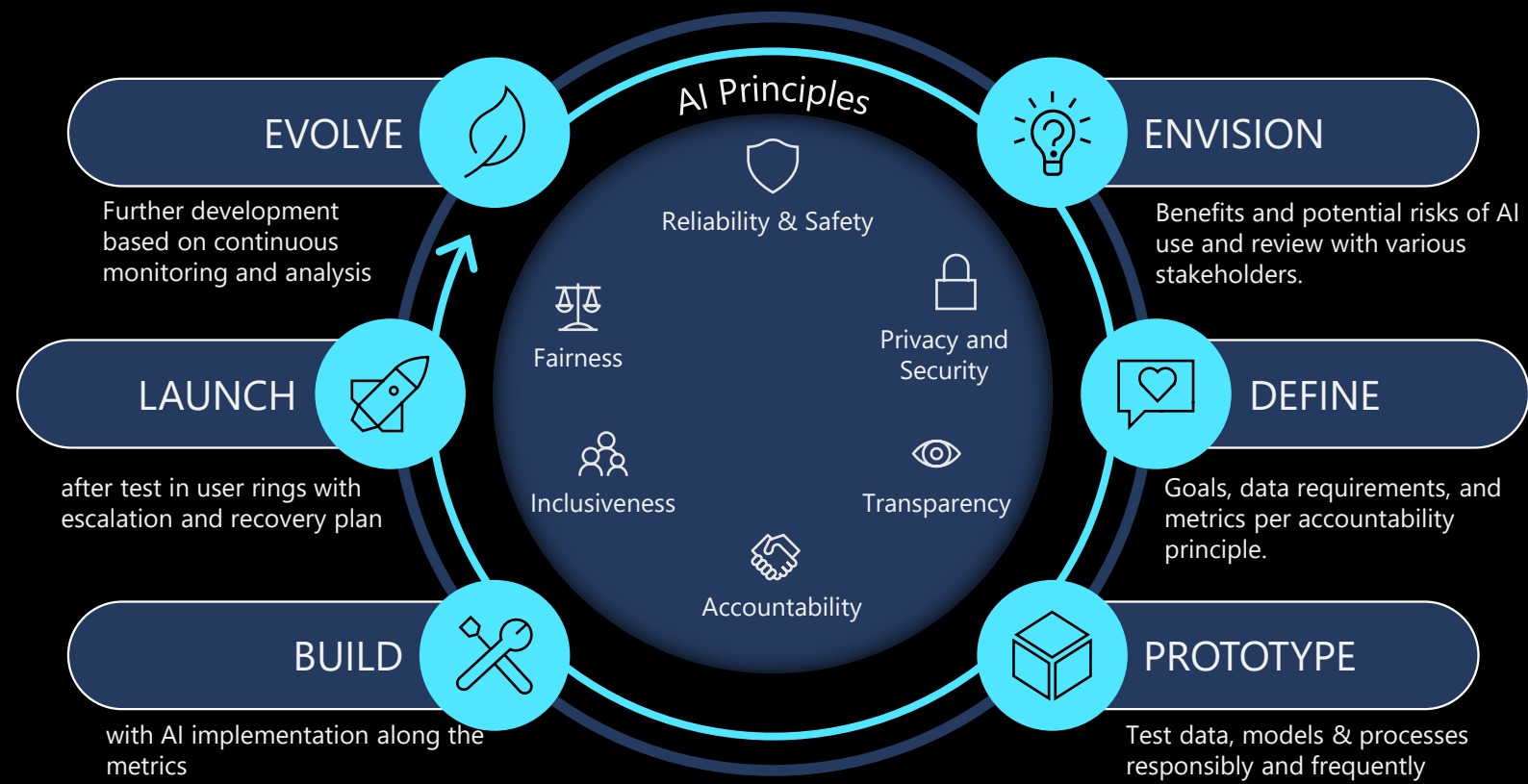


Tools





Microsoft's AI Practices



Putting AI principles into practice



Human-AI Guidelines

Conversational AI Guidelines

Inclusive Design Guidelines

AI Fairness Checklist

Datasheets for Datasets

<Insert Project Name>

RAI Champ: <insert RAI Champ Name>

TIMING

Timing expectations from customer and account team.

<insert text here>

MS ROLE

Explain MS role in the development, delivery, etc. of the solution.

<insert text here>

OVERVIEW

Customer

- Provide a high-level overview of the customer, industry, and pertinent information that may be relevant. Include any additional info on other projects happening with the customer that may be impacted.

Solution

- Provide detailed information on the technology being proposed, use cases, and additional information on the system.

Status Quo

Define processes the customer is using now in absence of this technology. What are they expecting the technology to improve.

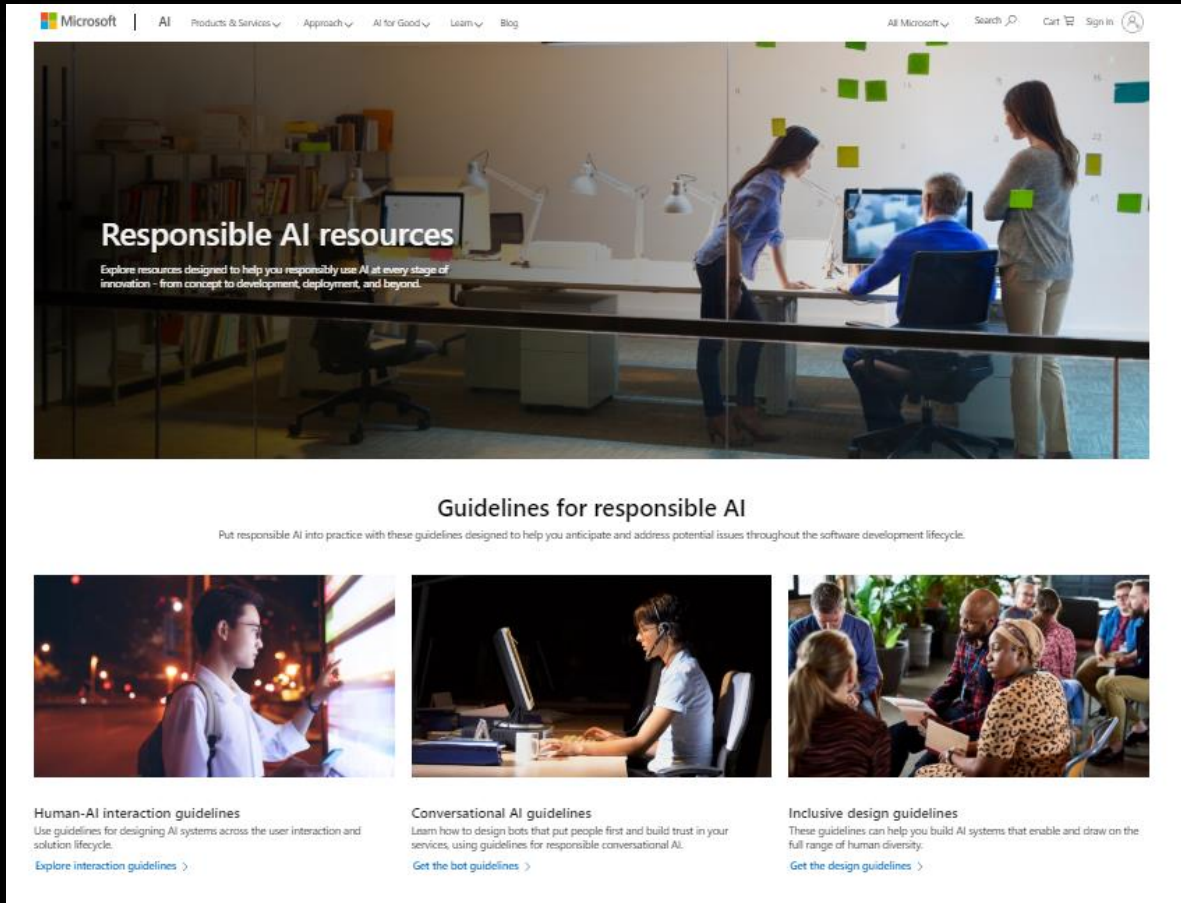
SENSITIVE USE SCENARIOS

- Denial of Consequential Services?
 - Include details relating to this sensitive use (if any, remove if not).
- Risk of Harm?
 - Include details relating to this sensitive use (if any, remove if not).
- Infringe on Human Rights?
 - Include details relating to this sensitive use (if any, remove if not).

RECOMMENDATION

- What are your recommendations for next steps following the review?
- Do we proceed with guidance? If so, what type of guidance is necessary?
- Does the customer need to do more assessment on the project prior to proceeding? If so, what work needs to be done? What is the timing?
- Would you recommend we not proceed, and if so, why?
- Do you believe no guidance is necessary? If so, why?


Responsible AI Resource Center




Responsible AI resources
Explore resources designed to help you responsibly use AI at every stage of innovation - from concept to development, deployment, and beyond.

Guidelines for responsible AI


Put responsible AI into practice with these guidelines designed to help you anticipate and address potential issues throughout the software development lifecycle.



Human-AI interaction guidelines
Use guidelines for designing AI systems across the user interaction and solution lifecycle.
[Explore interaction guidelines >](#)



Conversational AI guidelines
Learn how to design bots that put people first and build trust in your services, using guidelines for responsible conversational AI.
[Get the bot guidelines >](#)



Inclusive design guidelines
These guidelines can help you build AI systems that enable and draw on the full range of human diversity.
[Get the design guidelines >](#)

Centralized resource for practitioners to put responsible AI into action across the development lifecycle

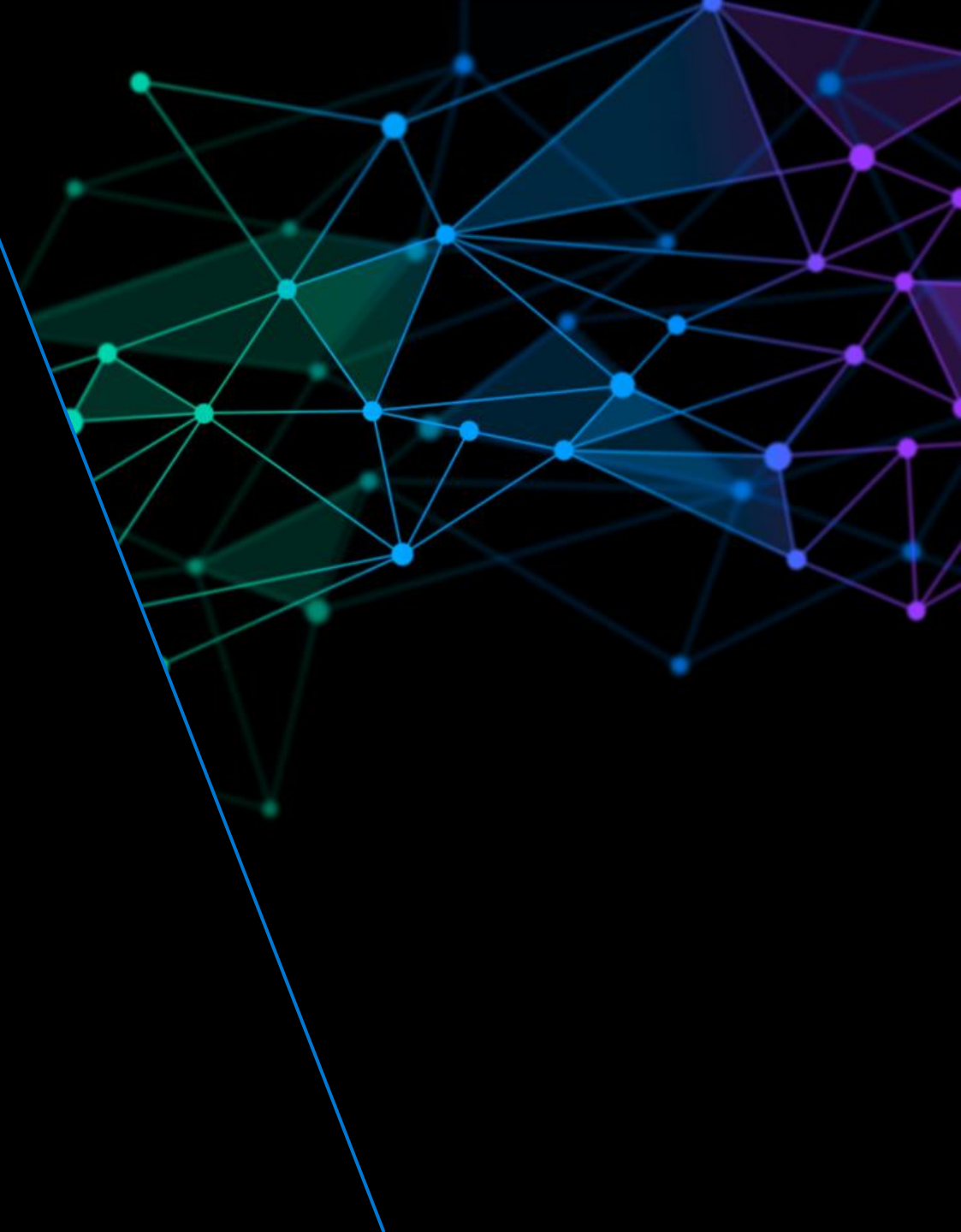
Guidelines and practices to help anticipate and address potential issues

Tooling innovation to help you understand, protect, and control AI models

Insights and perspectives from leading experts from across Microsoft

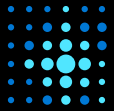
<https://aka.ms/rairesources>

Responsible AI Tools



Tools for responsible AI

Understand



Interpret
Machine Learning



FairLearn



Error Analysis

Protect



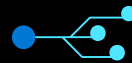
Homomorphic
Encryption



Differential
Privacy



Presidio



Confidential
Machine Learning

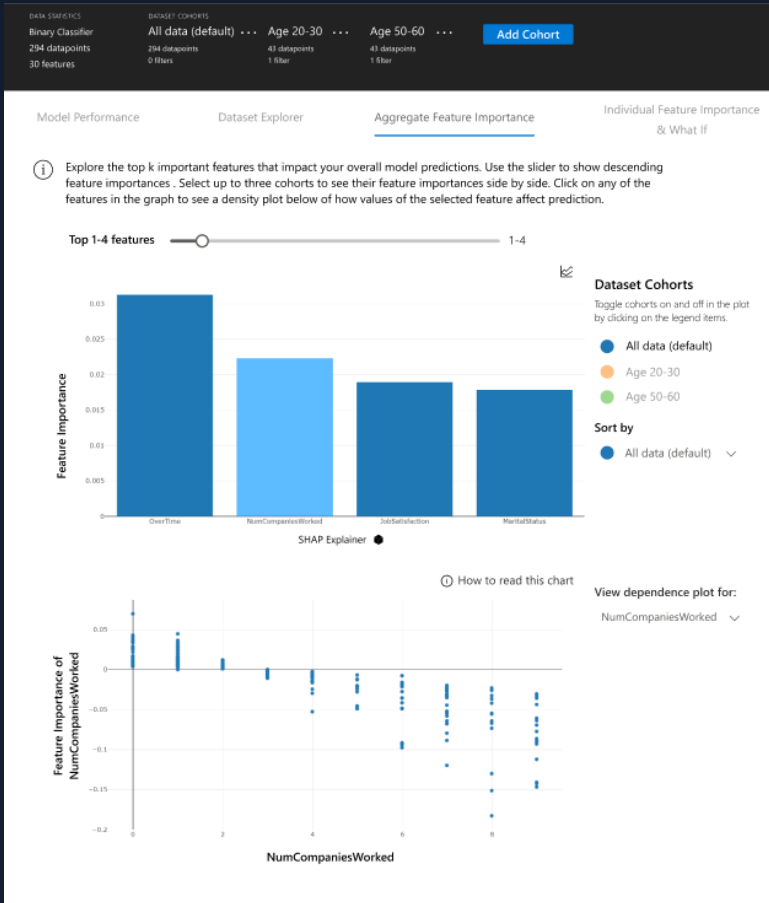
Control



MLOPs



Audit trail
Datasheets



Interpret

Glassbox and blackbox interpretability methods for tabular data



Interpret-community

Additional interpretability techniques for tabular data



Interpret-text

Interpretability methods for text data



DiCE

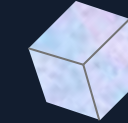
Diverse Counterfactual Explanations



Blackbox Models:

Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras,

Explainers: SHAP, LIME, Global Surrogate, Feature Permutation



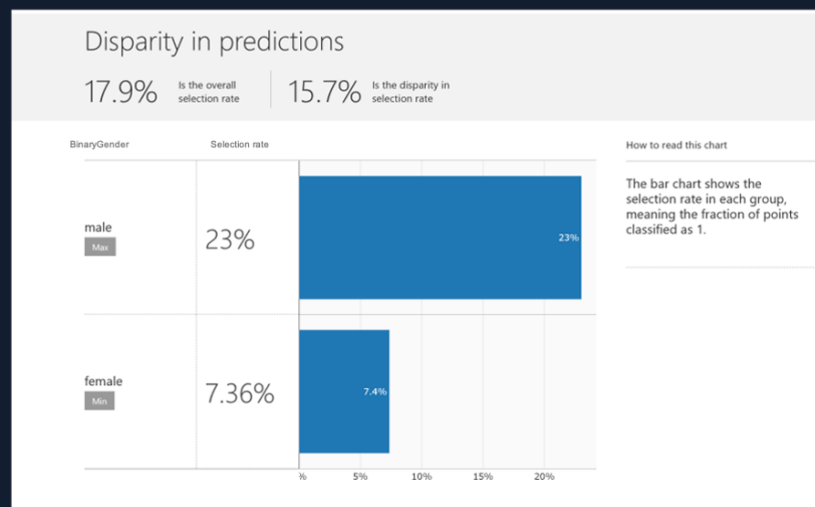
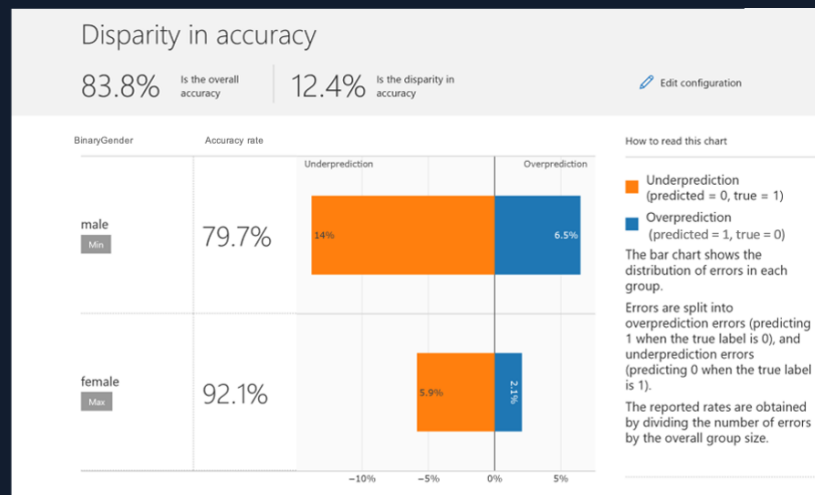
Glassbox Models:

Model Types: Linear Models, Decision Trees, Decision Rules, Explainable Boosting Machines



Azureml-interpret

AzureML SDK wrapper for Interpret and Interpret-community



Fairness Assessment:

Use common **fairness metrics** and an **interactive dashboard** to assess which groups of people may be negatively impacted.

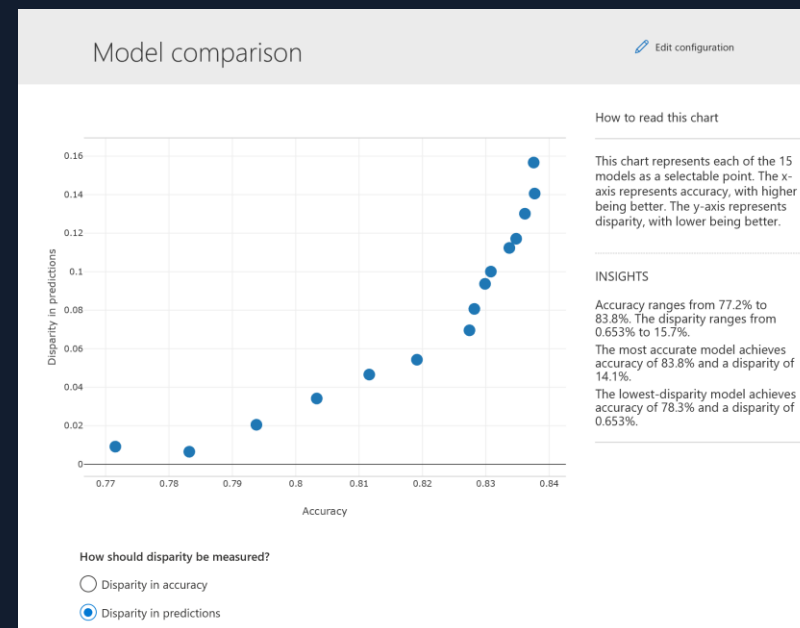
Model Formats: Python models using scikit predict convention, Scikit, Tensorflow, Pytorch, Keras

Metrics: 15+ Common group fairness metrics

Model Types: Classification, Regression

Fairness Mitigation:

Use state-of-the-art algorithms to mitigate unfairness in your classification and regression models.





"Anonymized data isn't"

Re-identification attacks exploit existing knowledge or data to reconstruct anonymized records.



Differential Privacy

Enables evaluations of machine learning while hiding the information contribution of individual data sets.



SmartNoise

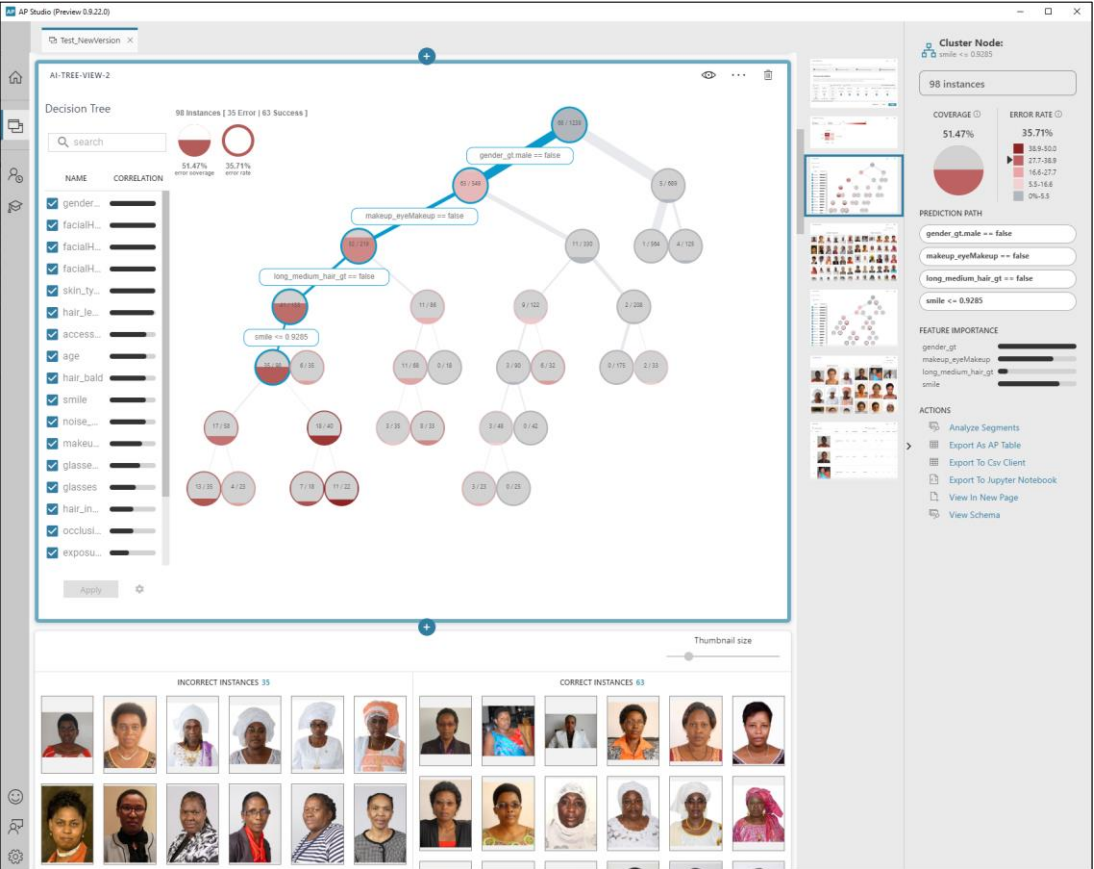
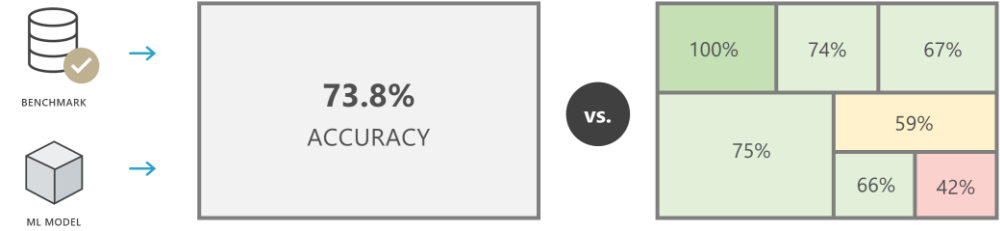
Open source implementation of the Differential Privacy Standard (Microsoft and Harvard University).





Error Analysis Toolkit

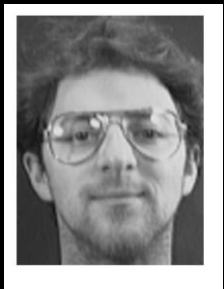
Error Analysis is a Responsible AI toolkit that enables you to get a deeper understanding of machine learning model errors. When evaluating a machine learning model, aggregate accuracy is not sufficient and single-score evaluation may hide important conditions of inaccuracies. Use Error Analysis to identify cohorts with higher error rates and diagnose the root causes behind these errors.



AI Security

AI Attacks

- Input Attacks
- Poisoning Attacks
- Adversarial Attacks
- Model Inversion Attacks



Training Image
[Fredrikson et al, 2015](#)



Reproduced Image

AI Quality Checkpoints

- Data and Model Inspections & Audits
(Fairlearn, Error Analysis Report, InterpretML etc.)
- Unattended Feedback-Loops
- Data Drift, Model Drift Audits
- Threat Modeling AI/ML Systems

<https://docs.microsoft.com/en-us/security/engineering/threat-modeling-ai/ml>

Food for Thought and Discussion

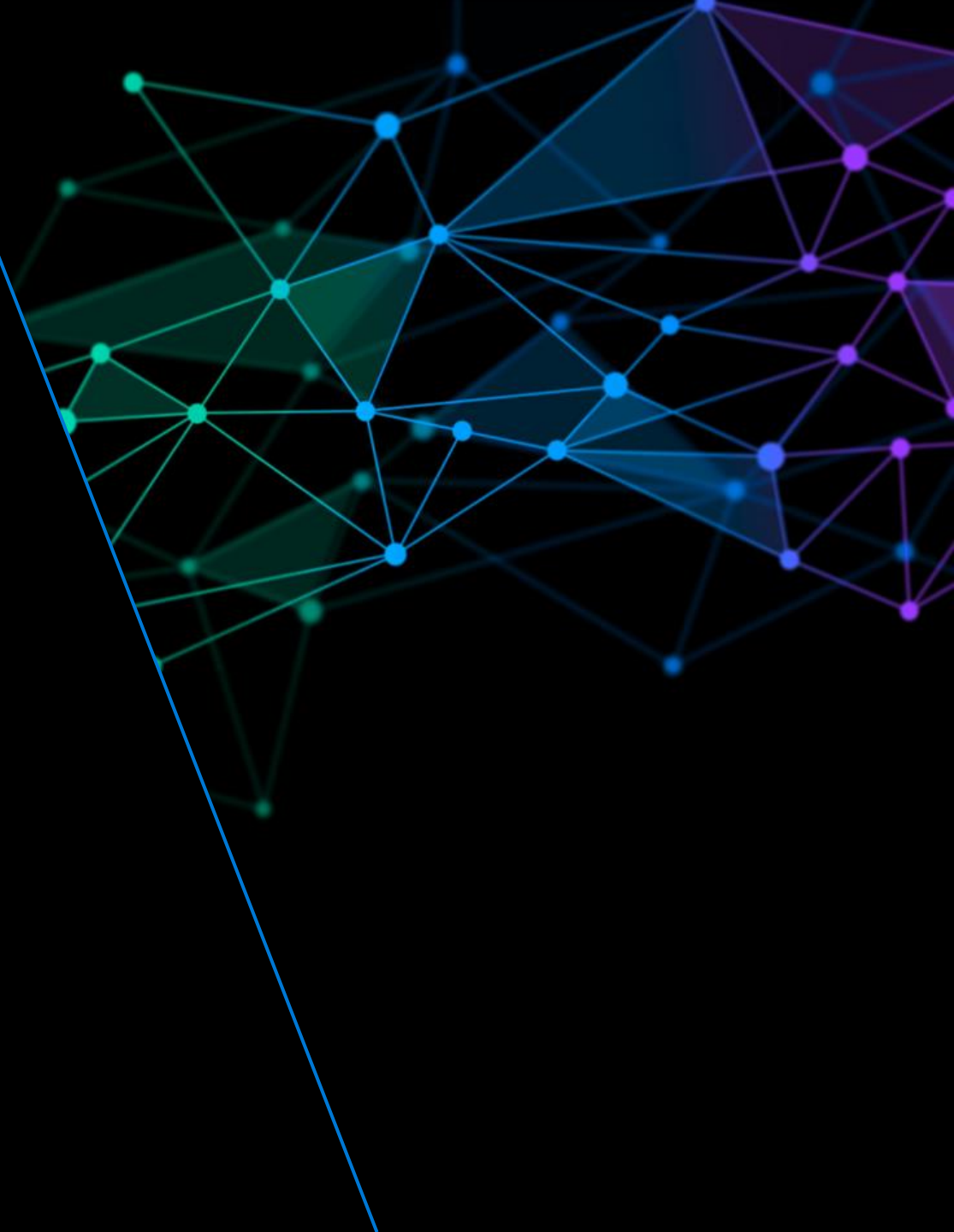
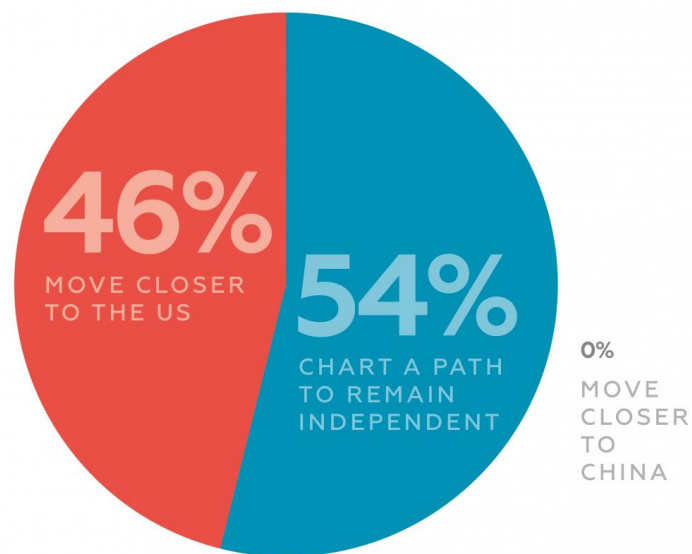


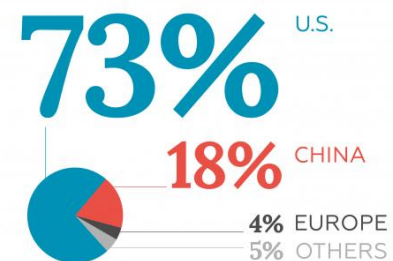
FIG. 8: HOW SHOULD THE EU POSITION ITSELF IN THE US-CHINA TECH CONFRONTATION?



Source: DGAP Stakeholder Survey 2021

DGAP

FIG. 1: MARKET CAPITALIZATION OF TOP 70 TECH COMPANIES



Source: "Tech regulation: The Brussels effect, continued.", The Economist (2020)

FIG. 2: TECHNOLOGICAL ADOPTION IN MANUFACTURING



Source: Debora Revoltella et al., "Adoption of digital technologies by firms in Europe and the US: Evidence from the EIB Investment Survey" (March 18, 2020)

DGAP

AI arms race

<https://www.weforum.org/agenda/2021/02/heres-what-you-need-to-know-about-the-new-ai-arms-race/>

Here's what you need to know about the new AI 'arms race'



Flying the flag: The United States has made substantial investments in AI Image: REUTERS/Yuri Gripas.

22 Feb 2021

Yori Kamphuis

AI Researcher, yorikamphuis.nl

Stefan Leijnen

Professor of AI, Utrecht University of Applied Sciences




- The US and China both outpace the EU on investment in AI.
- AI dominance can take on many forms.
- The EU could champion a citizen-driven approach to AI.

"Whoever becomes the leader in AI [or artificial intelligence] will become the ruler of the world," Vladimir Putin once [famously said](#).

<https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence>

Regulation on a European Approach for Artificial Intelligence



AI will impact all industries

TECH \ ARTIFICIAL INTELLIGENCE \

GitHub and OpenAI launch a new AI tool that generates its own code

Microsoft gets a taste of OpenAI's tech

By Dave Gershgorin | Jun 29, 2021, 1:46pm EDT

f t SHARE




Photo: GitHub

GitHub and OpenAI have launched a technical preview of a new AI tool called Copilot, which lives inside the Visual Studio Code editor and autocompletes code snippets.

GitHub Copilot

Sign up >

Technical Preview

Your AI pair programmer

With GitHub Copilot, get suggestions for whole lines or entire functions right inside your editor.

Sign up >

```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```

Replay

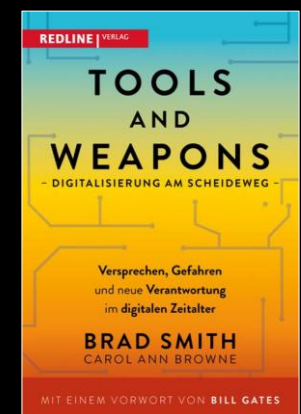
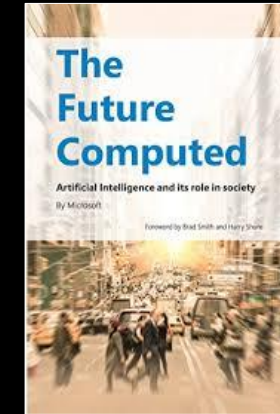
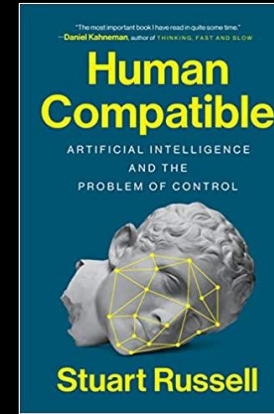
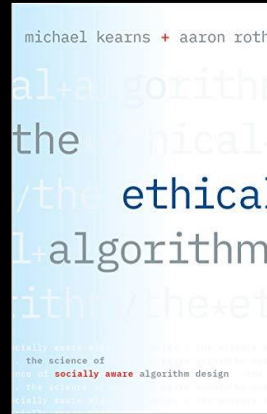
Powered by OpenAI

<https://copilot.github.com/>

Learn about Responsible AI

The screenshot shows the Microsoft Learn interface for a learning path titled "Identify principles and practices for responsible AI". The page includes a navigation bar with links to Docs, Documentation, Learn, Q&A, and Code Samples. The main content area features a hexagonal icon with a person and a magnifying glass, followed by the title and a description: "It is imperative to reflect on the implications of AI in business. In this learning path, you will be provided with guidelines to assist in setting up principles and a governance model in your organization. You will also be provided with resources, best practices, and tools." Below the description are tabs for "Intermediate", "Functional Consultant", "Business User", "Azure", "Dynamics 365", and "Microsoft 365". A "Prerequisites" section indicates "None". A "Save" button is at the bottom. The "Modules in this learning path" section lists two modules: "Identify guiding principles for responsible AI" (56 min, 9 Units, 4.7 rating) and "Identify governing practices for responsible AI" (35 min, 6 Units, 4.7 rating).

<https://docs.microsoft.com/en-us/learn/paths/responsible-ai-business-principles/>



Dokumentation

www.microsoft.com/ai/responsible-ai

azure.microsoft.com/blog/build-ai-you-can-trust-with-responsible-ml

docs.microsoft.com/azure/machine-learning/concept-responsible-ml

Vorstellung "The Ethical Algorithm": youtube.com/watch?v=cPZoP640ZhY

Responsible AI Werkzeuge

github.com/interpretml/interpret

github.com/fairlearn/fairlearn

github.com/opendifferentialprivacy

The background is a dark, monochromatic abstract composition. It features several layers of wavy, translucent lines that create a sense of depth and movement. Scattered throughout the scene are numerous small, dark, spherical particles, some of which appear to be in motion, creating a dynamic and textured effect. The overall aesthetic is modern and technological.

Thank You



AI is the New Normal

