Árnyék a gépezetben: mit kezdjünk a mesterséges intelligencia fekete dobozával?

Huszti Dániel daniel.huszti1@ibm.com

+36 20 823 5737

(c) 2023 IBM Corporation



Algorithmic bias isn't just unfair — it's bad for business

If it's not deployed wisely, artificial intelligence can turn consumers off.

By Kalinda Ukanwa Updated May 23, 2021, 3:00 a.m.

IDEAS

Data science during COVID-19: Some reassembly reauired

Most likely, the assumptions behind your data science model or the patterns in 018 / 4:04 PM / UPDATED 2 YEARS AGO vour data did not survive the coronavirus pandemic. Here's how to address the challenges of model drift

content creators They claim YouTube's algorithms discriminate against black users

YouTube sued for using AI to racially profile

Amazon scraps secret AI recruiting tool that showed bias against women



The Washington Post

Democracy Dies in Darkness

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Putting AI Governance in perspective

MLOps

- •Business glossary
- Policy management
- Metadata management
- Data lineage
- Classification
- Reference data
 management

Data Governance

- Data wrangling
- Mix-and-match coding, automated and visual model building
- Decision Optimization
- Dashboards
- •Jobs
- •Online/batch
- deployment
- Pipelines, CI/CD, APIs

- •Capture model metadata
- Accuracy, drift, bias monitoring
- Explainability
- Risk dashboards
- Regulatory compliance
- Risk management
- •Issue / alert management

AI Governance

Human Oversight

- Avoid undermining human autonomy
- Cost model integrating decision costs
- Performance Confidence intervals
- Optimized decisions allocation thresholds
- •Decisions allocation gain estimates

IBM Cloud Pak for Data

Regulation and reputation are the driving factors behind AI governance

English Search European Iome > ... > Contract rules > Digital contracts > Liability Rules for Artificial Intelligence European Commission Liability Rules for Artificial Intelligence The European approach to artificial intelligence (AI) will help build a esilient Europe for the Digital Decade where people and businesses can enjoy the benefits of Al. Home > Library > SOV.UK ✓ Gov artificial intelligence Home > Business and industry > Science and innovation > Artificial intelligence Establishing a pro-innovation approach to regulating Al Department fo Department fo Office for Artificial Business, Energy Digital, Culture, Intelligence & Industrial Strategy Media & Sport Policy paper artificial intelligence Establishing a pro-innovation approach to regulating Al Updated 20 July 2022

EN Q English Search Shaping Europe's digital future Menu Proposal for a Regulation laying down harmonised rules on POLICY AND LEGISLATION | Publication 21 April 2021 Proposal for a Regulation laving down harmonised rules on

engadget

UK police fail to use facial recognition ethically and legally, study finds

Steve Dent October 31, 2022, 9:20 am

Netherlands

Court of Audit

An Audit of 9 Algorithms used by the Dutch Government

Responsible use of algorithms by government agencies is possible but not always the case in practice. The Netherlands Court of Audit found that 3 out of 9 algorithms it audited met all the basic requirements, but the stimulation : from inadequate control over the algorithm's performance and impact to bias, data leaks and unauthorised access.

BlackRock shelves unexplainable AI liquidity models

Risk USA: Neural nets beat other models in tests, but results could not be explained



Amazon scraps secret AI recruiting tool that showed bias against women

Nearly all (97%) respondents believe that regulation will impact them to some extent and 95% believe that at least part of their business will be affected by the EU regulations specifically.

25% have yet to establish any meaningful Responsible AI capabilities.

Accenture - From AI compliance to competitive advantage, 2022

"Fewer than 20% of executives strongly agree that their organizations' practices and actions on AI ethics match (or exceed) their stated principles and values."

IBM and Oxford Economics - AI ethics in action, 2021

AI Governance solution

IBM solution is built on 3 pillars to meet clients on their maturity curve

Lifecycle governance

Monitor, catalog, and govern AI models from anywhere, throughout the AI lifecycle

Risk management

Manage risk and compliance to business standards, through automated facts and workflow management

IBM's differentiation

Comprehensive: only vendor to support all three layers of AI governance

Automated: automated facts collection and lineage tracking within python notebooks

Open: AI vendor-neutral, supports governance of models built and deployed in third party tools

Unified: unified cataloging for models & data and unified notions of data quality for AI and reporting

Regulations-driven: support for AI regulations natively

Active policy enforcement: support for policies and rules to automate regulatory compliance

Regulatory compliance external A

Ensure clients adhere to external AI regulations for audit and compliance

(c) 2023 IBM Corporation

In practice, this means that...



Business stakeholders need to ... (C-Suite, CPO/CRO, Business Owner)

- ✓ Define new responsibilities, policies, guidelines and processes based on AI principles & enterprise values
- ✓ Establish organizational structures
- ✓ Understand regulation and translate to business & technical requirements
- Raise awareness and train leaders and practitioners
- ✓ Assess maturity, risk level and conformity
- $\checkmark\,$ Oversee, manage and mitigate risks
- ✓ Design architecture, pipelines and governance rules for trust and scale



Technical stakeholders need to ... (Data Scientist, Ops Lead, Risk Manager)

- ✓ Adhere to data governance rules
- ✓ Keep track of security risks
- ✓ Document every new and updated model
- ✓ Find and fix direct and indirect bias in each data set and model
- ✓ Find and fix model accuracy drift & data consistency drift in each model
- ✓ Explain model decisions as requested
- ✓ Find and fix adversarial attacks on your data and models
- ✓ Complete all model review process activities



Which is easier said than done



Putting AI Governance to work

A typical customer context



Technology View – AI Governance Cloud Pak for Data components in scope for Trusted AI



Bias detection

Continuous calculation of model fairness

- Analyze deployed model predictions for bias
- Collect and aggregate bias data for dashboards and alerts
- Find non-feature data correlations
- Use a corrected model for "de-biased" predictions

Ensuring fairness in model scoring

BLE



Model Explainability

Explain model predictions

- Show the most influential features
- Explain in natural language
- Available API for prediction explanations

What-if analysis

- Experiment with values
- Assess effects of changes to features

Understand model outcomes



Explain		Inspect										
Transaction e5a5ee83c	ce3c1b52148	16b447211d5cd3-1	Received on Jul 07, 2021, 5:21:59		oyed model gage Default - Pro	oduction Pr	Language roduction Not applicable					
Predicted DENII Confidenc 94.00	ce level		three features influenc	- Production mo ing the model's	predicted outcor	me are Loc	ce that the outcome of this transactio ation, Income, and SalePrice. The top e model toward a predicted outcome	three features AppliedOnline,				
For this tra	insaction, eac		el has been assigned a per ure influenced the model t				gly the feature has influenced the model's	predicted outcome. A negative				
4	33	3.0%		Other	features			Positive Influence				
3	0%						Explain Inspect					
TIVE WEIG	0%	8.9	%	5.5%		Fo by of	Reaching a different predicted or the model to have predicted a different y more than the minimum value may affe t changing the prediction.	outcome for this transaction, the valu				
	.0%		%	5.5%	-4.9%	Fo by of	or the model to have predicted a different y more than the minimum value may affer f changing the prediction.	outcome for this transaction, the valu				s indicate a greater likeliho
	.0%		5.9%	5.5%	-4.9%	Fo by of	or the model to have predicted a different y more than the minimum value may affe f changing the prediction.	outcome for this transaction, the valu t the minimum change of other featur	res for the model to predict a differen		her feature importance numbe	s indicate a greater likeliho
-1	0%	8.9	8.9%	5.5%	-4.9%	Fo by of	or the model to have predicted a different y more than the minimum value may affer t changing the prediction. Analyze controllable features only A Feature	outcome for this transaction, the valu t the minimum change of other featur Original value	res for the model to predict a different of the model to predict a different of the second seco	nt outcome. Hig	her feature importance numbe	s indicate a greater likelihoo
-1	0%	8.9	5.9%	5.5%	-4.9%	Fo by of	or the model to have predicted a different y more than the minimum value may affer t changing the prediction. Analyze controllable features only A Feature Location	outcome for this transaction, the value t the minimum change of other featur Original value	New value	nt outcome. Hig	her feature importance numbe Value for a different ou L100	() Importance 1.00
-1	0%	5.9 ation Inco		5.5%		Foo by of	or the model to have predicted a different y more than the minimum value may affer t changing the prediction.	outcome for this transaction, the value t the minimum change of other featur Original value L120 50445	New value L120 50445	v	Value for a different ou L100 50445	indicate a greater likelihoo Importance 1.00 0.00
-1	0%			Number_of_Car.		Foo by of	or the model to have predicted a different y more than the minimum value may affer to hanging the prediction.	outcome for this transaction, the value t the minimum change of other featur Original value L120 50445 NO	New value 120 58445 NO		Value for a different ou 50445 NO	Importance 0.00 0.00
-1	0%			Number_of_Car.	Residence	Foo by of	or the model to have predicted a different more than the minimum value may after changing the prediction.	utcome for this transaction, the value the minimum change of other feature of triginal value L120 50445 NO Public Housing	New value L120 50445 NO Public Housing		Value for a different ou 100 50445 NO Public Housing	Importance 1.00 0.00 0.00 0.00
-1	0%			Number_of_Car.	Residence	Foo by of	or the model to have predicted a different more than the minimum value may affe changing the prediction.	outcome for this transaction, the value the minimum change of other feature of the second sec	New value L120 50445 NO Public Housing 10		Value for a different ou L00 50445 NO Public Housing 10	Importance 1.00 0.00 0.00 0.00 0.00 0.00
-1	0%			Number_of_Car.	Residence	Foo by of	or the model to have predicted a different more than the minimum value may affe changing the prediction. Analyze controllable features only Feature Location Location Income AppliedOnline Residence Vrs_at_Current_Address Vrs_with_Current_Employee	outcome for this transaction, the value the minimum change of other feature clipped clipped clipped clipped clipped clipped clipped cl	New value Value L20 S0445 NO Public Housing 10 8		Value for a different ou L100 50445 NO Public Housing 10 8	indicate a greater likelihoo Importance 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
-1	0%			Number_of_Car.	Residence	Foo by of	or the model to have predicted a different ymore than the minimum value may affer tehanging the prediction.	outcome for this transaction, the value the minimum change of other feature of original value L120 50445 NO Public Housing 10 8 8 1	New value Ilian Ilian <thilian< th=""> Ilian Ilian</thilian<>		Value for a different ou 2400 50445 NO Public Housing 10 8	Importance 100 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00

Drift detection

Measure the degree to which a model has moved away from reality

- Drop in accuracy reality has changed, as shown by the scoring data
- Drop in consistency reality is the same, the events vary

Drift monitoring and alerts

 Degradation of model performance can trigger retraining and redeployment
 (c) 2023 IBM Corporation Handle changing scenarios





Manage risk across the enterprise – Open Pages

- Consistent holistic views of risk and compliance
- Drive GRC adoption
- Embedded selfservice reporting, analytics, and dashboarding



How to get started? Launch a pilot project on AI Governance



Recently announced at IBM Think 2023 Put AI to work with **watsonx**

Scale and accelerate the impact of AI with trusted data.



Enable fine-tuned models to be managed through market leading governance and lifecycle management capabilities

