

Summary

Machine Learning – Tools and applications for policy – Lecture 10

Iman van Lelyveld – Michiel Nijhuis

DNB Data Science Hub



Summary

1. Discuss some things that can go wrong

- survivorship bias, input errors and deceit
- fairness and discrimination

2. What is the reaction of authorities?

Summary

- What can go wrong?
- Why is this sensitive?
- Should we intervene?
- The exam: what to expect?
- Lessons learned

- No clips for this lecture

What could possibly go wrong?

- We have discussed **overfitting** to your training data
 - **Solutions:** more data, regularization
- We generally assume that the **Data Generating Process (DGP)** is the same between training and production
 - In practice this might not always be true:
 - The training set was put together with bespoke (==expensive) information and only some features were retained. The real data is from a different population
 - We might have missed crucial feature information
 - The world changes ...

0. Training data
1. Model maintenance
2. Survivorship bias: WWII airplanes
3. Crucial input error: Skin cancer (Esteva et al. (2017))
4. Derailing a DNN
5. Neutrality of the methods
6. Discrimination: assessing recidivism (Wu and Zhang (2016) and Wu and Zhang (2017))

0. Challenges of Real Data Collection

- Collection of real data can be challenging because:
 - High cost and resource requirements
 - Limited availability of data sources
 - Ethical or privacy concerns
- Alternative could be to generate synthetic data
- Benefits of Synthetic Data
 - Prototyping: Useful for prototyping and proof of concept.
 - Resource Savings: Cost-effective alternative to real data collection.
 - Privacy: Can be generated without sensitive information.
 - Model Development: development and testing in the absence of real data.

- May not perfectly represent the real-world distribution.
- Model performance on synthetic data may not directly translate to real data.
- Quality of synthetic data depends on generative model and source data.
- Careful validation, rigorous testing, and ethical considerations are essential.

1. Model maintenance —Complex Models Erode Boundaries

- Dependable software enforces strict abstraction boundaries. Data is data. Model is model. etc
- Unfortunately, it is difficult to enforce strict abstraction boundaries for machine learning systems by prescribing specific intended behavior (Sculley et al. (2015)).
- Resulting problems:
 - **Entanglement:** Machine learning systems mix signals together, entangling them and making isolation of improvements impossible.
 - ▶ **CACE principle:** Changing Anything Changes Everything
 - ▶ Mitigation strategies: isolate models and serve ensembles or detecting changes in prediction behavior as they occur.
 - **Correction Cascades:** There are often situations in which models for problem A exists, but a solution for a slightly different problem A' is required.
 - **Undeclared Consumers:** Without access controls, some consumers of results may silently use the output of a given model as an input to another system.

- Unstable Data Dependencies
- Underutilized Data Dependencies: Similar to packages that are mostly unneeded, underutilized data dependencies are input signals that provide little incremental modeling benefit.
 - Legacy features
 - Bundled Features
 - ϵ -Features
 - Correlated features

- **Direct**

- What if the model leads to the selection of a particular trading strategy, this influences prices, influencing the trading strategy
 - Is this bad? Maybe points to retraining or modelling the feedback loop explicitly

- **Hidden**

- What if Firm A has an algorithm setting price as the average price at $t - 1$ and Firm B has the rule $p_{t-1}^A + 1 \dots$

Feedback Loops

11

- Direct

- What if the model leads to the selection of a particular trading strategy, this influences prices, influencing the trading strategy
- Is this bad? Maybe points to retraining or modelling the feedback loop explicitly

- Hidden

- What if Firm A has an algorithm setting price as the average price at $t - 1$ and Firm B has the rule $p_{t-1}^A + 1 \dots$

Price starts at 1.7 million ...

The screenshot shows an Amazon search results page for the book "The Making of a Fly: The Genetics of Animal Design" by Peter A. Lawrence. The results are filtered to show "New" items. Two editions are listed:

- New** 1-2 of 2 offers
Price + Shipping Condition Seller Information Buying Options
\$1,730,045.91 New Seller: profmath Add to Cart
+ \$3.99 shipping Seller rating: ***** 93% positive over the past 12 months.
(8,393 total ratings)
In Stock. Ships from NJ, United States.
[Domestic shipping rates and return policy](#)
Brand new. Perfect condition. Satisfaction Guaranteed.
- New** \$2,198,177.95 New Seller: bordesbook Add to Cart
+ \$3.99 shipping Seller rating: ***** 93% positive over the past 12 months.
(125,001 total ratings)
In Stock. Ships from United States.
[Domestic shipping rates and return policy](#)
New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!

A sidebar on the right provides a "Price at a Glance" summary:

Price	at a Glance
Unit	\$30.00
Used:	from \$35.54
New:	\$1,730,045.91

Other buttons in the sidebar include "Safe Online Shopping" and "Buy It Now".

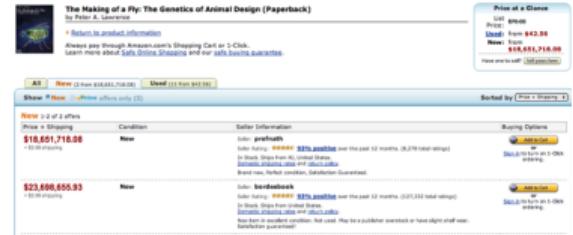
- Direct

- What if the model leads to the selection of a particular trading strategy, this influences prices, influencing the trading strategy
- Is this bad? Maybe points to retraining or modelling the feedback loop explicitly

- Hidden

- What if Firm A has an algorithm setting price as the average price at $t - 1$ and Firm B has the rule $p_{t-1}^A + 1 \dots$

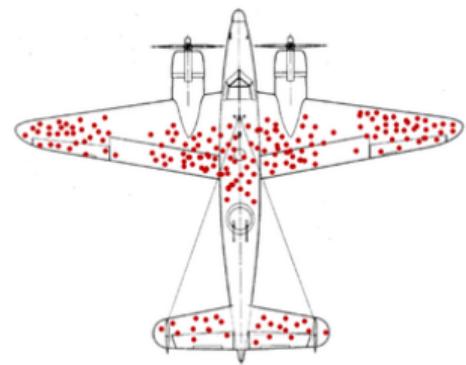
... and jumps to 18 million



- **Glue Code:** code to get to standard packages can be more work than writing things yourself. On the other hand we have the "Iwannadoitmyself" bug
 - Solution: wrap black-box packages into common API's
- **Pipeline Jungles**
- **Dead Experimental Codepaths**
 - Knight Capital's system losing \$ 465 million in 45 minutes, apparently because of unexpected behavior from obsolete experimental codepaths
- **Abstraction Debt.** The above issues highlight the fact that there is a distinct lack of strong abstractions to support ML systems (check out [MLFlow](#))
- **Smells**
 - Plain-Old-Data Type Smell
 - Multiple-Language Smell.
 - Prototype Smell.

2. Survivorship bias

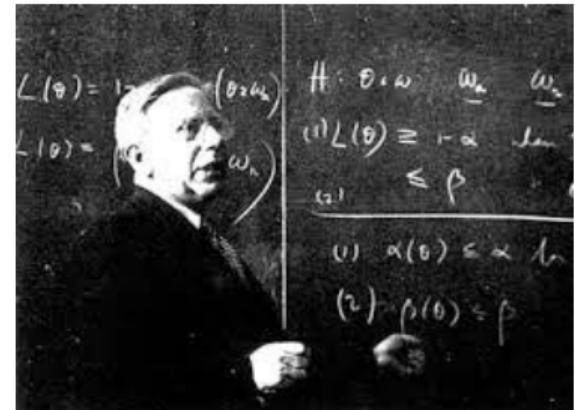
- In WWII the Royal Airforce wanted to find out how to increase aircraft resilience
- Red dots indicate damage of returning bombers
- Where would you add armor?



Source: [Wikipedia](#)

2. Survivorship bias

- In WWII the Royal Airforce wanted to find out how to increase aircraft resilience
 - Red dots indicate damage of returning bombers
 - Where would you add armor?
-
- The army wanted to add armor where they observed damage
 - Luckily statistician Abraham Wald noted that the sample suffered from survivorship bias: sample was only the surviving airplanes
 - planes hit in engines or cockpit do not return



3. Skin cancer

- Skin cancer, the most common human malignancy, is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by more (invasive) tests
- Esteva et al. (2017, Nature) train a CNN using a dataset of 129,450 clinical images consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images
- It would be great if this method could be rolled out to your smartphone ...



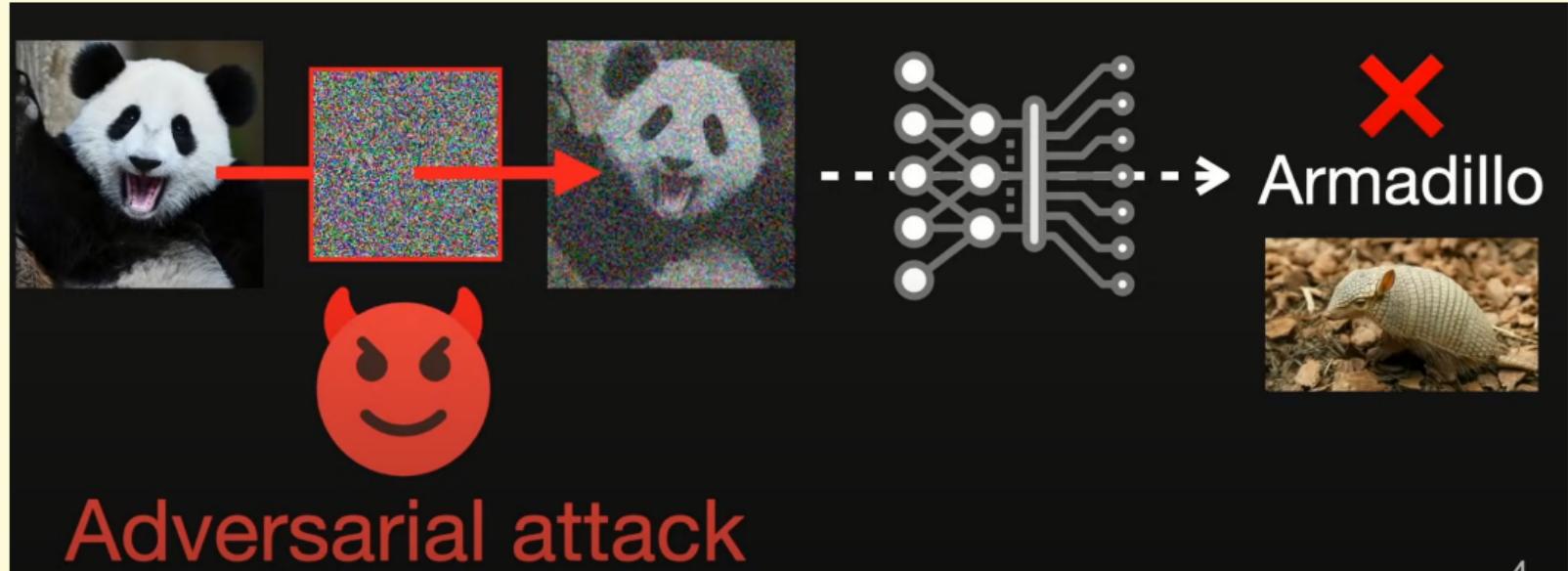
3. Skin cancer

- Skin cancer, the most common human malignancy, is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by more (invasive) tests
- Esteva et al. (2017, Nature) train a CNN using a dataset of 129,450 clinical images consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images
- It would be great if this method could be rolled out to your smartphone ...
- ... but turns out that a ruler is bad for you



4. Derailing a Deep Neural Network

15



Source: Das *et al*, 2020, [Github repo](#)

An attack can derail a Deep Neural Network

16

Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

A Control Sidebar

ADVERSARIAL ATTACK
PGD
Strength: 0.05

FILTER GRAPH
 Show full graph
 Show pinned only
 Show highlighted only

HIGHLIGHT PATHWAYS
Highlight pathways most excited by attack.
Neurons: top 45 % in each layer
Connections: top 50 %

B Graph Summary View

GIANT PANDA BOTH ARMADILLO EXPLOITED BY ATTACK

C Detail View

mixed4d-46

Med. Activ. — Giant panda — Armadillo

Attack strength

Feature Vis Examples from data

brown bird

46

5. ML is not neutral

- We've discussed many ways in which ML models can be flawed
- One more way is that ML models allow for non-linearity (and dislike (relatively) high variance)
- Fuster et al. (2022): guidance to identify the specific groups most likely to win or lose from the change in technology.
 - Consider the decision of a lender who uses a single exogenous variable (e.g., a borrower characteristic such as income) to predict default
 - Who wins or loses with new technology depends on 1) the **functional form of the new technology**, and 2) differences in the **distribution of income** across groups
 - **Primitive prediction** technology which simply uses the mean level of a single characteristic to predict default → the predicted default rate = same for all borrowers
 - More sophisticated **linear model** with default linearly decreasing → groups with lower income than the mean will lose
 - **Convex quadratic function**: penalize groups with higher variance of the characteristic
 - etc. etc.

What is a fair definitions of fair?

18

The screenshot shows the homepage of the AI Fairness 360 toolkit. At the top, there's a navigation bar with links for Home, Demo, Resources, Events, Videos, and Community. The 'Home' link is highlighted with a blue underline. Below the navigation is a section titled 'AI Fairness 360' with a subtext: 'This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.' To the right of this text is a circular icon containing a blue scale symbol. Below this section are four buttons: 'Python API Docs', 'Get Python Code', and 'Get R Code'. A note below the buttons says 'Not sure what to do first? Start here!'. The main content area is divided into several sections: 'Read More' (with a link to 'Fairness and bias mitigation concepts, terminology, and tools'), 'Try a Web Demo' (with a link to 'Check through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.'), 'Watch Videos' (with a link to 'Watch videos to learn more about AI Fairness 360.'), 'Read a paper' (with a link to 'Read a paper describing how we designed AI fairness 360.'), 'Use Tutorials' (with a link to 'Step through a set of in-depth examples that introduce you to how to code that checks and mitigates bias in different industry and application domains.'), 'Ask a Question' (with a link to 'Join our AI F360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.'), 'View Notebooks' (with a link to 'Open a directory of Jupyter Notebooks in GitHub that provide examples of bias detection and mitigation in simple datasets. Then share your own notebooks.'), and 'Contribute' (with a link to 'You can add new metrics and algorithms in GitHub. Share Jupyter notebooks showing how you have examined and mitigated bias in your machine learning application.').

Source: [IBM AIF 360](#). Also see [Godata driven blog](#).

Why is this relevant for finance?

19

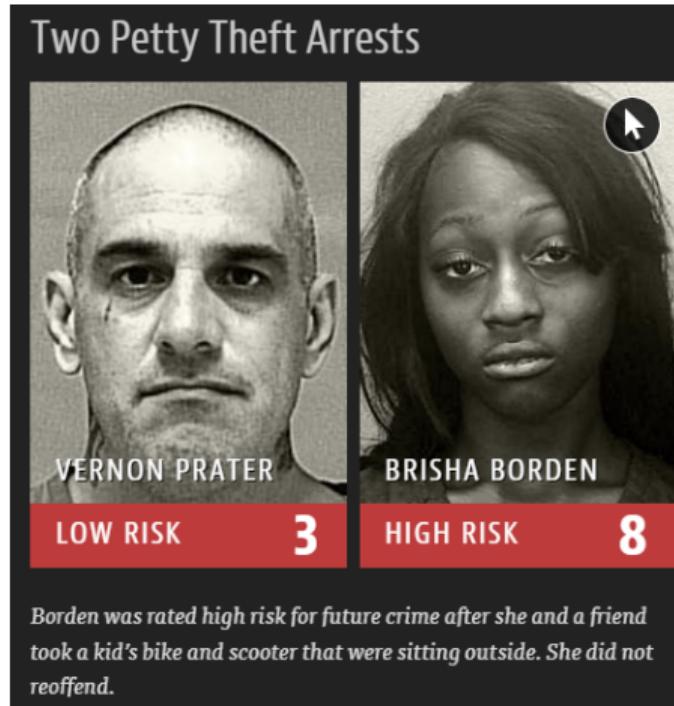
- Regardless of what you think of positive discrimination, these concepts can be useful in finance
 - What will be the performance of my portfolio if I can't trade in particular sin stocks or sanctioned regions
 - How similar are two vectors or matrices
 - etc etc

6. Do we discriminate in who to send to jail?

20

- Similar to the (Ludwig and Mullainathan 2023), the Compass system is a commercial product to help judges make decisions about incarceration.
- [Pro Publica article](#) discusses how this system seems biased
-

How Compass scores get it wrong



Source: Pro Publica

How Compass scores get it wrong

21



Source: Pro Publica

How Compass scores get it wrong



Source: Pro Publica

How Compass scores get it wrong

21



Source: Pro Publica

How Compass scores get it wrong



Source: Pro Publica

How Compass scores get it wrong

21



Source: Pro Publica

How Compass scores get it wrong



Source: [Pro Publica](#)

- Wu and Zhang (2016) used ML to classify 'criminal' faces. They compare criminal ID photo's with website scraped ones
- This led to a storm of protest – what issues can you think of?



(a) Three samples in criminal ID photo set S_c .



(b) Three samples in non-criminal ID photo set S_n

- Wu and Zhang (2016) used ML to classify 'criminal' faces. They compare criminal ID photo's with website scraped ones
- This led to a storm of protest – **what issues can you think of?**
- Critics point to:
 - sample size was too small
 - non-criminals are smiling (according to westerners) and have a white collar
 - claim of causality
- In Wu and Zhang (2017) the authors discuss (refute?) each of these point



(a) Three samples in criminal ID photo set S_c .



(b) Three samples in non-criminal ID photo set S_n

Summary

What can go wrong?

Why is this sensitive?

Should we intervene?

The exam: what to expect?

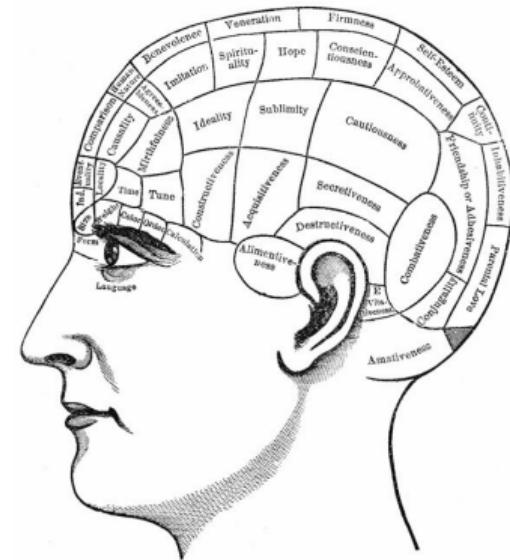
Lessons learned



Why is this an important and touchy subject?

24

- In 19th century many believed in **phrenology** – the shape of the skull determines behaviour
- This is now widely discredited



Source: [Wikipedia](#)

Why is this an important and touchy subject?

24

- In 19th century many believed in **phrenology** –
the shape of the skull determines behaviour
- This is now widely discredited
- But how does a police officer decide to check a suspect?



Why is this an important and touchy subject?

24

- In 19th century many believed in **phrenology** –
the shape of the skull determines behaviour
- This is now widely discredited
- But how does a police officer decide to check a
suspect?
- How are loans and mortgages granted?
- Who is responsible for a market crash caused by
algorithmic trading?
- What if ML is used for Anti Money Laundering?
And what if crooks use an adversarial neural
network to counter this (cf Triepels (2019))

- The insurer must have systems and processes in place that prevent the implementation of AI applications that lead to **discriminatory outcomes**
- When setting up such systems and processes for AI applications, questions are:
 - How to organise the process to challenge possible discriminatory bias in input variables?
 - How can results be tested for discriminatory bias? Possibilities are **adversarial modeling**, as well as two identical test groups, where only one discriminatory (proxy) variable differs between groups. If the model differs significantly, this may indicate potential discriminatory bias in the model. Hypothetical cases can also be used
 - How can testing for discriminatory bias become more refined and robust? For example, specific testing of biases in the false positive outcomes (instead of just testing the overall model outcomes)
- If for an AI model it cannot be sufficiently established that it does not contain unauthorized discriminatory biases, should you use it in processes that directly affect customers (for example pricing and acceptance, fraud detection)?

Summary

What can go wrong?

Why is this sensitive?

Should we intervene?

The exam: what to expect?

Lessons learned

Why Google thinks we need to regulate AI

Companies cannot just build new technology and let market forces decide how it will be used

SUNDAR PICHAI

+ Add to myFT

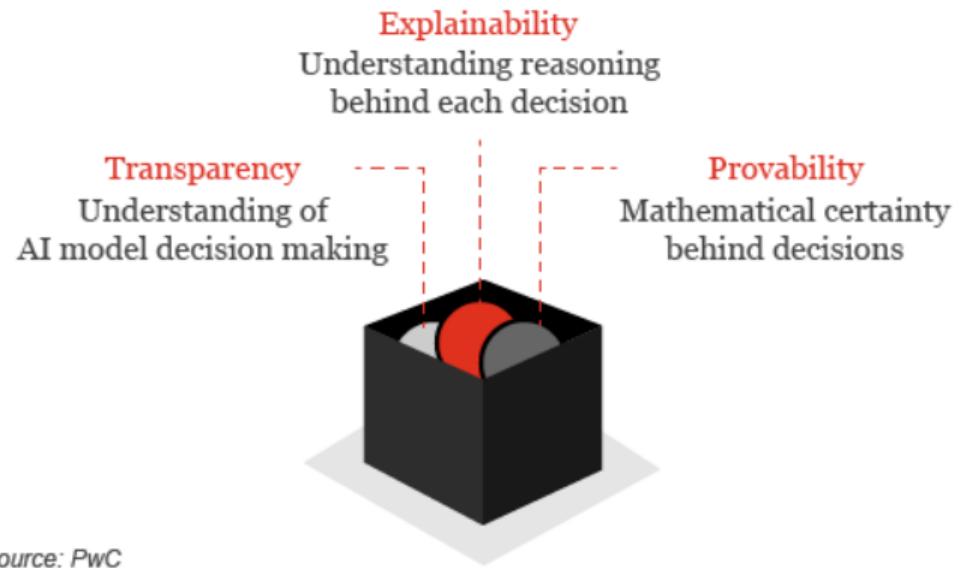


DataScience
Hub

Further words of caution

28

- Explainability



Source: PwC

Further words of caution

28

- Explainability
- Environmental impact



- Explainability
- Environmental impact

Common carbon footprint benchmarks

in lbs of CO₂ equivalent

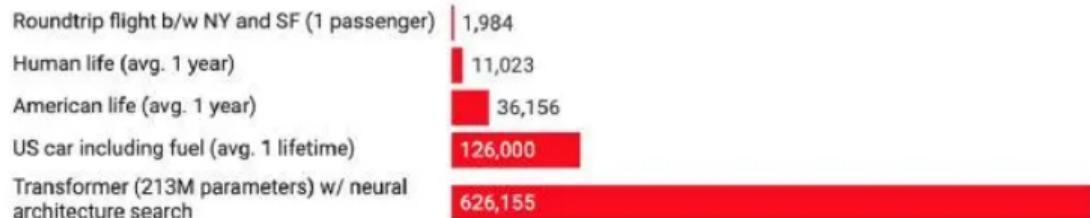


Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Further words of caution

- Explainability
- Environmental impact
- Market power (cf Farboodi et al. (2022))



- Surveys: Broeders et al. (2018), Bundy (2017), Chakraborty and Joseph (2017), Frost (2020), Petralia et al. (2019), and World Economic Forum (2018)
- Regulator sandboxes
 - Bank regulation can be over the top for fintech startups
 - Several supervisors provide 'bank light'-regimes
 - iForum
- Fintech hubs
 - BIS, Singapore, Lithuania
- Guidelines (van der Burgt (2019))
 - SAFEST: Soundness, Accountability, Fairness, Ethics, Skills, Transparency
- Proposed EU regulation



EUROPEAN COMMISSION

Brussels, 21.4.2021

COM(2021) 206 final

2021/0106(COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

EXPLANATORY MEMORANDUM

1. CONTEXT OF THE PROPOSAL

1.1. Reasons for and objectives of the proposal

This explanatory memorandum accompanies the proposal for a Regulation laying down harmonised rules on artificial

- On 21 April 2021, the European Commission published its Proposal for a Regulation laying down harmonised rules on artificial intelligence ([Artificial Intelligence Regulation \(AIR\)](#))

- On 21 April 2021, the European Commission published its Proposal for a Regulation laying down harmonised rules on artificial intelligence ([Artificial Intelligence Regulation \(AIR\)](#))
- **AI defined** as “a software that is developed with one or more of approaches and technique and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with”

- On 21 April 2021, the European Commission published its Proposal for a Regulation laying down harmonised rules on artificial intelligence ([Artificial Intelligence Regulation \(AIR\)](#))
- **AI defined** as “a software that is developed with one or more of approaches and technique and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with”
- **Aim:** The regulatory proposal aims to provide AI developers, deployers and users with clear requirements and obligations regarding specific uses of AI.

- On 21 April 2021, the European Commission published its Proposal for a Regulation laying down harmonised rules on artificial intelligence ([Artificial Intelligence Regulation \(AIR\)](#))
- **AI defined** as “a software that is developed with one or more of approaches and technique and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with”
- **Aim:** The regulatory proposal aims to provide AI developers, deployers and users with clear requirements and obligations regarding specific uses of AI.
- **Scope:** This regulation applies to a natural or legal person, public authority, agency or other body that:
 - developed (or commissioned) an AI system to bring to market (directly or indirectly) ([AI systems providers](#))
 - is using an AI system under its authority ([AI systems users](#))
 - is established in a third country if the system output is used in the Union.

The regulation follows a risk-based approach

32

- **Unacceptable risk:** All AI systems considered as a clear threat to the safety, livelihoods and rights of people will be banned (e.g. **social scoring by governments**, toys using voice assistance that encourages dangerous behaviour.)

¹see Annex III, Article 5 (b) of the Proposal

The regulation follows a risk-based approach

32

- **Unacceptable risk:** All AI systems considered as a clear threat to the safety, livelihoods and rights of people will be banned (e.g. [social scoring by governments](#), toys using voice assistance that encourages dangerous behaviour.)
- **High risk:** All high-risk AI systems will be subject to “strict obligations” before they can be put on the market. As one of the high risk applications include systems that determine [individuals' credit scores](#).¹

¹see Annex III, Article 5 (b) of the Proposal

- **Unacceptable risk:** All AI systems considered as a clear threat to the safety, livelihoods and rights of people will be banned (e.g. [social scoring by governments](#), toys using voice assistance that encourages dangerous behaviour.)
- **High risk:** All high-risk AI systems will be subject to “strict obligations” before they can be put on the market. As one of the high risk applications include systems that determine [individuals' credit scores](#).¹
- **Limited risk:** Limited risk AI systems, such as [chatbots](#), will be subject to transparency obligations, with users clearly told that they are interacting with intelligent software rather than a human, and voluntary codes of conduct.

¹see Annex III, Article 5 (b) of the Proposal

The regulation follows a risk-based approach

- **Unacceptable risk:** All AI systems considered as a clear threat to the safety, livelihoods and rights of people will be banned (e.g. [social scoring by governments](#), toys using voice assistance that encourages dangerous behaviour.)
- **High risk:** All high-risk AI systems will be subject to “strict obligations” before they can be put on the market. As one of the high risk applications include systems that determine [individuals' credit scores](#).¹
- **Limited risk:** Limited risk AI systems, such as [chatbots](#), will be subject to transparency obligations, with users clearly told that they are interacting with intelligent software rather than a human, and voluntary codes of conduct.
- **Minimal risk:** free use of applications such as AI-enabled [video games](#) or [spam filters](#). The vast majority of AI systems currently used in the EU fall into this category

¹see Annex III, Article 5 (b) of the Proposal

- AI system to comply with requirements on i) data and data governance, ii) technical documentation, iii) record-keeping, iv) human oversight and v) accuracy, robustness and cybersecurity Art. 8 to 15
- Put a quality management system in place Art. 17(3) → Art. 74 CRD
- Establish risk management procedure Art. 9(9) → Art. 74 of CRD
- Technical documentation is part of internal governance framework
Art. 18(2) → Art. 74 of CRD
- Maintain automatic logs of high-risk AI systems Art. 20 (2) → Art. 74 CRD
- Carry out **conformity assessment** Art. 19 and 43 which shall be carried out as **part of SREP**
Art. 43(2) → Art. 97 to 101 of CRD)
- Establish and document a post-market monitoring system Art. 61(1) + (4)
- Obtain EU declaration of conformity and CE marking Art. 48(1) + 49

- Ensure human oversight measures Art.29 (2)
- Ensure that input data is relevant in view of the intended purpose of AI system
Art. 29 (3)
- Monitor the AI system Art. 29(4) → Art. 74 of CRD
- Maintain automatic logs of high-risk AI systems Art. 29(5) → Art. 74 of CRD

Summary

- What can go wrong?
- Why is this sensitive?
- Should we intervene?
- The exam: what to expect?**
- Lessons learned

1. If a feature contributes to a good prediction performance of customer default, we should use it to decline bad risks.
2. What happens if you train a Perceptron on a linearly inseparable set? Why is this the case? What can be done about it?
3. A colleague presents a model with 99% accuracy – Given your experience in the assignments, what would be your reaction?
4. What is the bias-variance trade-off? How does regularization influence this trade-off?
5. What is a hyperplane? How is it put to use in an SVM?
6. A colleague notes that the correlation of X and Y is very high. Will this help you to predict Y ? Explain.
7. How could you employ unsupervised learning in a supervised classification problem?

8. Does a radial basis function always outperform the default, linear basis function in a SVM?
9. What type of neural network would you use for a linearly separable problem?
10. What makes it difficult to apply ML in finance?
11. When do we standardize or normalize features? Provide an example for when it matters and an example for when it doesn't matter
12. What do we optimize in a regression problem versus a classification problem?
13. Is it possible to evaluate the performance of a classification problem with the MSE only?
14. Sklearn has an option of DecisionTreeClassifiers called `min_impurity_split`, a float parameter, which sets the "Threshold for early stopping in tree growth". When would you change this parameter?

What is the take-home exam going to be like?

38

- The same setup as Assignment 2
 - We provide a data set, you have to read it in, clean it, describe it
 - You can choose features and engineer them
 - Then you get a number of questions about training, validating and predictions
- The topic is going to be **sentiment** and their **effect on price formation**
- We have upload a take-home exam
- In comparison to this one, we will put more emphasis on written explanations
- We will release the exam on **14th of December**, 17.00 AM.
 - Please post your clarifying questions beforehand in the discussion thread before 10.00 AM Friday 15 December. We can then prepare the Q&A.
- We will have a Q&A on Zoom on **15th of December**, 15.00 PM.

- The main assessment form is a take-home exam at the end of the course. Make sure you allocate enough time before the deadline. Note that it is in the exam week!
- The total number of points attainable for the take-home exam is 70
- The take-home exam is **individual**. Given that coding is inherently reusable and that discussing techniques is commendable, we realise that it can be difficult to discern unacceptable copying and smart and efficient reuse of code. We nevertheless hope that you will approach the take-home exam as your own work (instead of a class wide exercise)
- The grading for the take-home assignment is detailed in Table 1. Further guidance will follow in due course.
- Note that the assignments not necessarily lead into the take-home exam.

- The grade of the assignments follows roughly the same relative point allocation as the take-home exam.
- In addition to the take-home exam there will be two graded programming assignments with a maximum of 15 points attainable on each assignment.
- The two graded assignments should be submitted by self-selected groups of 3 students. Each assignment should be submitted as a Jupyter notebook on Canvas including necessary supporting documents
- The final grade is calculated as follows:
$$\text{Final grade} = (\text{Assignment 1} + \text{Assignment 2} + \text{Take-home exam})/10$$
- To pass the course you need to have a final grade of at least 5.5 (out of 10) as well as at least a 35 points for the take-home exam

Summary

- What can go wrong?
- Why is this sensitive?
- Should we intervene?
- The exam: what to expect?
- Lessons learned

Observations

- 2/3 lectures were (still) too short → allows for more material: deeper or broader?
- ...

Questions

- Assignments on the right level?
- What did you think of the guest lecture?

Survey

- Could you please take a couple of minutes to fill in the survey
- The link is below and you should have received an email invite
- <https://portal.evalytics.nl/evaluations/142003>

In this lecture we covered:

1. discussed aspects that can invalidate ML results – survivorship bias, input errors and sensitive results
2. introduced the regulatory response
3. discussed the parameters of the exam

Soundness

Accountability

Fairness

Ethics

Skills

Transparency

Soundness

Accountability

Fairness

Ethics

Skills

Transparency

- 1) *Ensure general compliance with regulatory obligations regarding AI applications.*
- 2) *Mitigate financial (and other relevant prudential) risks in the development and use of AI applications.*
- 3) *Pay special attention to the mitigation of model risk for material AI applications.*
- 4) *Safeguard and improve the quality of data used by AI applications.*
- 5) *Be in control of (the correct functioning of) procured and/or outsourced AI applications.*

Soundness

Accountability

Fairness

Ethics

Skills

Transparency

- 6) *Assign final accountability for AI applications and the management of associated risks clearly at the board of directors level.*
- 7) *Integrate accountability in the organisation's risk management framework.*
- 8) *Operationalise accountability with regard to external stakeholders.*

Soundness

Accountability

Fairness

Ethics

Skills

Transparency

- 9) *Define and operationalise the concept of fairness in relation to your AI applications.*
- 10) *Review (the outcomes of) AI applications for unintentional bias.*

Soundness

Accountability

Fairness

Ethics

Skills

Transparency

- 11) *Specify objectives, standards, and requirements in an ethical code, to guide the adoption and application of AI.*
- 12) *Align the (outcome of) AI applications with your organisation's legal obligations, values and principles.*

Soundness

Accountability

Fairness

Ethics

Skills

Transparency

- 13) *Ensure that senior management has a suitable understanding of AI (in relation to their roles and responsibilities).*
- 14) *Train risk management and compliance personnel in AI.*
- 15) *Develop awareness and understanding of AI within your organisation.*

Soundness

Accountability

Fairness

Ethics

Skills

Transparency

- 
- 16) *Be transparent about your policy and decisions regarding the adoption and use of AI internally.*
 - 17) *Advance traceability and explainability of AI driven decisions and model outcomes.*

-  Autoriteit Financiële Markt & De Nederlandsche Bank. (2019). Artificiële Intelligentie in de verzekeringssector – een verkenning (tech. rep.).
-  Broeders, D., Bruinshoofd, A., & Kilinç, M. (2018). Netwerken hebben invloed op besluitvorming door pensioenfondsen. 103(4758), 2–5.
-  Bundy, A. (2017). Preparing for the future of Artificial Intelligence [ISBN: 0013-9157]. AI & Society, 32(2), 285–287.
-  Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks. Bank of England Working Paper, 674. Retrieved December 7, 2018, from www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx
-  Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks [Publisher: Nature Publishing Group]. Nature, 542(7639), 115–118.
-  Farboodi, M., Matray, A., Veldkamp, L., & Venkateswaran, V. (2022). Where Has All the Data Gone? Review of Financial Studies, 35(7), 3101–3138.

-  Frost, J. (2020). The Economic Forces Driving FinTech Adoption across Countries. [DNB Working Paper, 663.](#)
-  Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. [Journal of Finance, 77\(1\), 5–47.](#)
-  Ludwig, J., & Mullainathan, S. (2023). Machine Learning as a Tool for Hypothesis Generation.
-  Petralia, K., Philippon, T., Rice, T., & Véron, N. (2019). [Banking disrupted?: Financial intermediation in an era of transformational technology](#) (Vol. 2019) [Publication Title: Geneva Reports on the World Economy Issue: 22 ISSN: 16078616].

- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *Advances in Neural Information Processing Systems*, 28. Retrieved November 10, 2022, from <https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcf2674f757a2463eba-Abstract.html>
- Triepels, R. (2019). *Anomaly Detection in the Shipping and Banking Industry* [Doctoral dissertation].
- van der Burgt, J. (2019). General principles for the use of Artificial Intelligence in the financial sector.
- World Economic Forum. (2018). The New Physics of Financial Services. Retrieved February 12, 2020, from www.deloitte.com/about
- Wu, X., & Zhang, X. (2016). Machine Learning of Criminality Perceptions. *arXiv*, 1611.04135.

-  Wu, X., & Zhang, X. (2017). Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135) [arXiv: 1611.04135]. [arXiv, 1611.04135](http://arxiv.org/abs/1611.04135). <http://arxiv.org/abs/1611.04135>