Strategic IT management - 37E00200

Automated decision making and ”explainable AI”

Esko Penttinen
Associate Professor, Information Systems Science, Aalto University School of Business
Chairman, XBRL Finland
Director, Real-Time Economy Competence Center

1

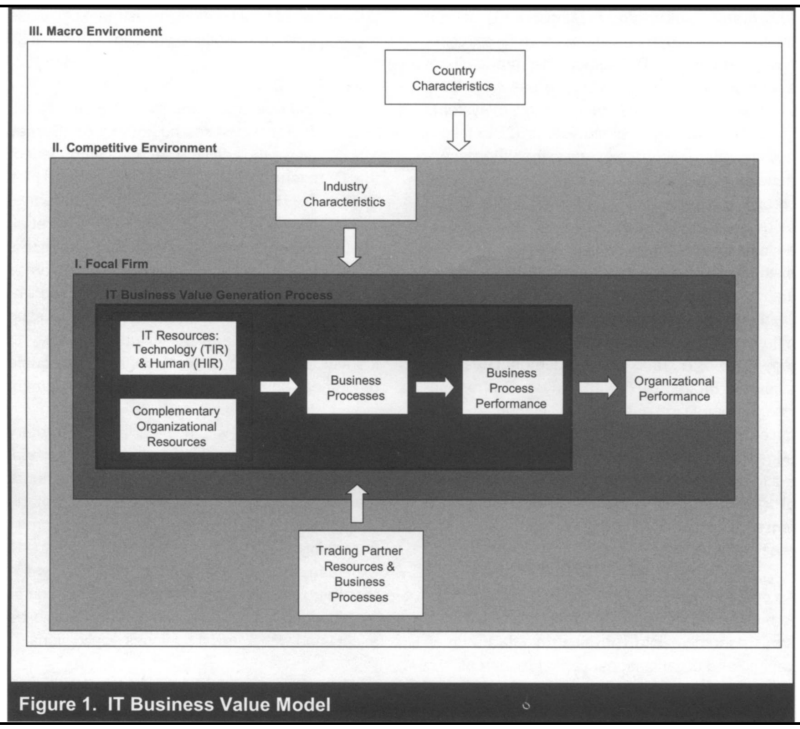


Figure 1. IT Business Value Model

Melville, Nigel, Kenneth Kraemer, and Vijay Gurbaxani. "Information technology and organizational performance: An integrative model of IT business value." *MIS quarterly* 28.2 (2004): 283-322.

2

Case Parlamentary ombudsman

- The tax administration sends approximately 300,000 reminder letters each year due to missing declarations, and more than 112,000 estimated tax decisions are made in automated processing. In these, the information system has completed all the stages of case processing and decision-making, without any natural person having participated in the processing of the case. This automatic assessment tax is also set at a rate of 25 percent tax increase on the estimated tax amount. Likewise, in corporate income taxation, estimated taxation decisions are made in automation and a five percent increase of the estimated tax amount is imposed. The control of the payment of taxes and the collection of taxes also take place in accordance with the settings made in the system in the automation system, other than for cases transferred to case-by-case collection. In the tax administration, the processing of the taxpayer's entire tax matter, including the consultation, decision-making and collection phases, can thus take place in automation without any natural person having participated in the processing of the tax matter.

Parlamentary ombudsman of Finland, 26.11.2019
https://www.oikeusasiamies.fi/r/fi/ratkaisut/-/eoar/3379/2018

Verohallinnon automatisoitu päätöksentekomenettely ei täytä perustuslain vaatimuksia - Tiedotteet

Hallinnonala

Edunvalvonta
Kieliasiat
Kirkko
Lapsen oikeudet
Liikenne ja viestintä
Muut
Opetus ja sivistys
Poliisi
Rikosseuraamusala
Sosiaalihuolto
Sosiaalivakuutus
Sotilasasiat ja puolustushallinto
Syyttäjälaitos
Terveydenhuolto
Tulli

Tiedotteet

Tiedotteita julkaistaan oikeusasiamiehen ratkaisuista, jotka ovat johtaneet toimenpiteeseen tai joilla voi muuten olla yleistä mielenkiintoa.

Verohallinnon automatisoitu päätöksentekomenettely ei täytä perustuslain vaatimuksia

Koska Verohallinnon automatisoitu verotus- ja päätöksentekomenettely ei perustu asianmukaiseen ja täsmälliseen lainsäädäntöön, jossa olisi otettu huomioon hyvän hallinnon ja oikeusturvan sekä virkavastuun asianmukainen toteutuminen, AOA piti sitä lainvastaisena.

Perustuslakivaliokunta on kiinnittänyt valtioneuvoston huomiota siihen, että automatisoituun päätöksentekomenettelyyn liittyy useita säätelemättömiä kysymyksiä. Valiokunnan mukaan sääntelytarpeesta

3

GDPR article 22 "Automated individual decision-making, including profiling"

Article 22
EU GDPR
"Automated individual decision-making, including profiling"

=> Recital: 71, 72
=> administrative fine: Art. 83 (5) lit b
=> Dossier: Automated Decision In Individual Cases, Profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
=> Article: 4

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
=> Dossier: Legitimate Interests (Data Subject), Opening Clause

(c) is based on the data subject's explicit consent.
=> Dossier: Consent

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller **shall implement suitable measures** to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
=> Recital: 70
=> Dossier: Legitimate Interests (Data Subject), Obligation

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.
=> Dossier: Legitimate Interests (Data Subject)

5

RPA (Robotic Process Automation) a prime example of lightweight IT

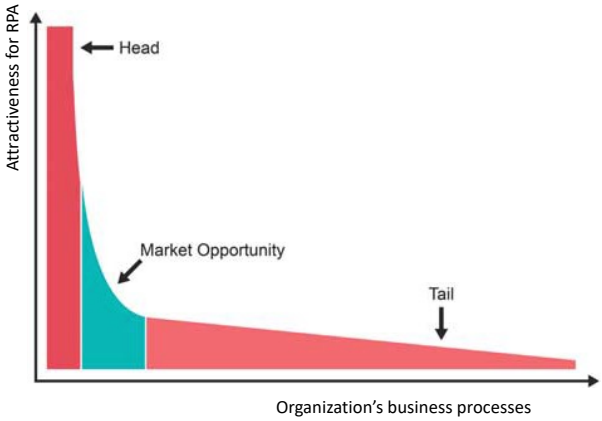
Table 1 Heavyweight and lightweight IT

	Heavyweight IT	*Lightweight IT*
	A knowledge regime, driven by IT professionals, enabled by systematic specification and proven digital technology and realized through software engineering	A knowledge regime, driven by competent users' need for solutions, enabled by the consumerisation of digital technology and realized through innovation processes
Profile	Back-end: Supporting documentation of work	Front-end: Supporting work processes
Owner	IT department	Users and vendors
Systems	Transaction systems	Process support, apps, BI
Technology	PCs, servers, databases, integration technology	Tablets, electronic whiteboards, mobile phones
IT architecture	Fully integrated solutions, centralised or distributed	Non-invasive solutions, frequently meshworks (heterogeneous networks)
Development culture	Systematics, quality, security	Innovation, experimentation
Problems	Increasing complexity, rising costs	Isolated gadgets, security
Discourse	Software engineering	Business and practice innovation

Bygstad, Bendik. "Generative innovation: a comparison of lightweight and heavyweight IT." *Journal of Information Technology* 32.2 (2017): 180-193.

6

Positioning RPA as an automation tool



Criteria for evaluating processes
High volume of transactions
Need to access multiple systems
Stable environment
Low cognitive requirements
Easy decomposition into unambiguous rules
Proneness to human error
Limited need for exception handling
Clear understanding of the current manual costs

Asatiani, A. & Penttinen, E. (2016) Turning Robotic Process Automation into Commercial Success – Case OpusCapita, *Journal of Information Technology Teaching Cases*, 6 (2), pp. 67-74.

7

RPA to handle sensitive data with RPA?

- Asatiani, A., Hakkarainen, T., Paaso, K. & Penttinen, E. (2023) Security by envelopment – A novel approach to data-security-oriented configuration of lightweight-automation systems. *European Journal of Information Systems*, forthcoming
- Asatiani et al. 2023b in “additional reading” folder in MyCourses

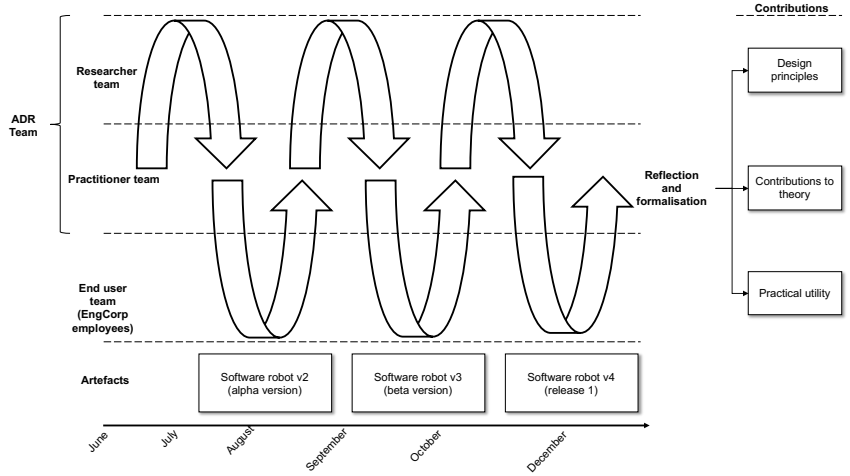
8

Case study RPA to report EU posted worker notifications

- Wärtsilä is a Finnish publicly listed industrial company
- Mandated by the EU Posted Workers Directive, Wärtsilä must compile and submit a report each time a Wärtsilä employee travels to another EU country for work
- The notification report is extensive and places administrative burden on Wärtsilä. The following systems need to be accessed to collect the data:
  - CRM : 34 items
  - SAP: 10 items
  - SAP HR: 17 items
- Most of these data are confidential and sensitive
  - Thus we ask: “How can Wärtsilä configure RPA to compile the report?”

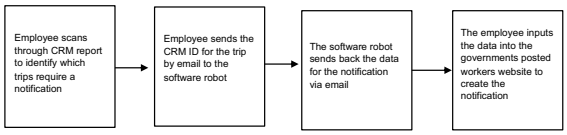
9

Action design research at Wärtsilä



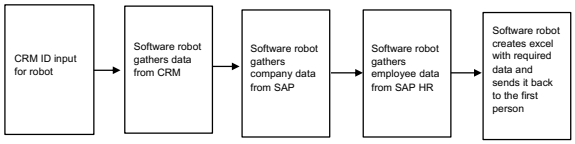
10

Process from human operator's perspective



11

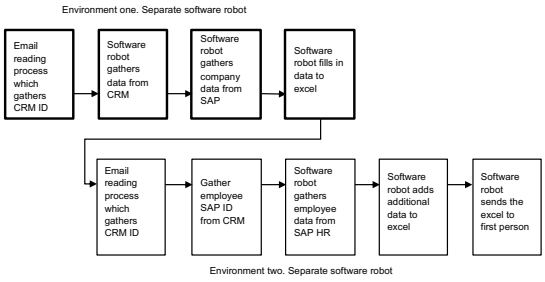
RPA V1 (prior to ADR process)



Challenges identified: (1) robot would need to handle SAP HR data in separate environment, and (2) anybody who knows the e-mail address of the software robot could trigger it by sending a CRM ID

12

RPA alpha cycle

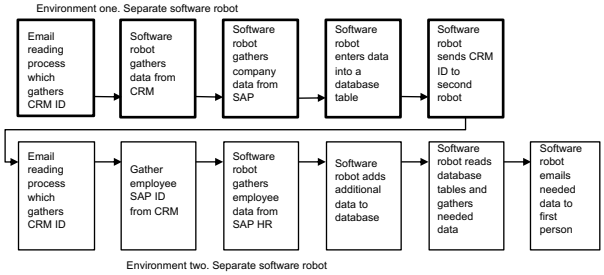


Intervention: create two separate software robots to handle SAP HR data and other data (SAP, CRM)

Challenges identified: efficiency (software robot required three minutes to operate) and database security issues emerged (Excel stored locally posed security threats)

13

RPA beta cycle

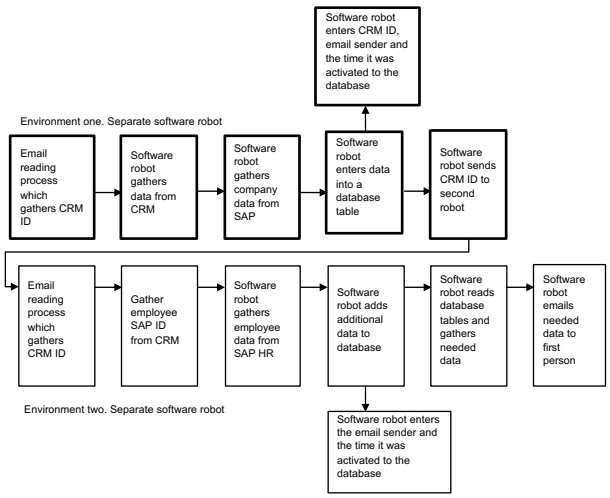


Intervention: no more Excel files, storing data in database and fetching data from there to the e-mail

Challenges identified: inability to monitor and audit access to data

14

RPA release

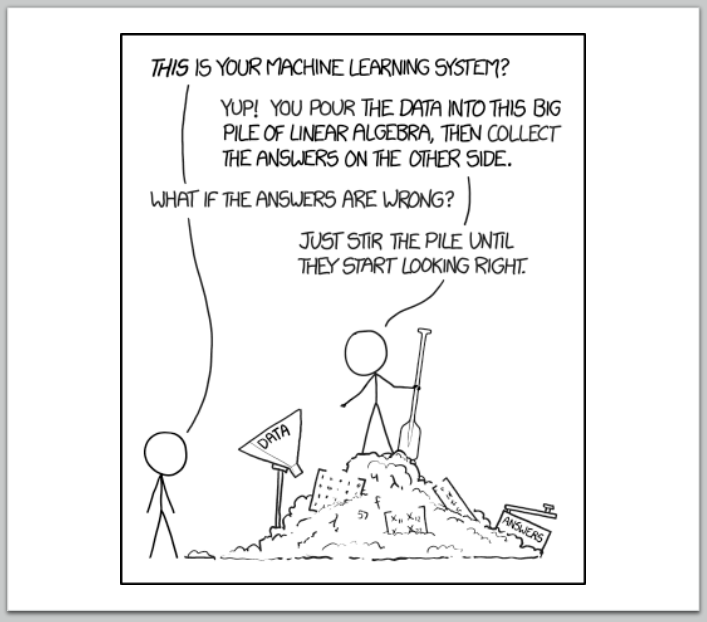


Intervention: audit trail for both environments established

15

Explainable AI

- What is explainable AI?
- Reasons for explainable AI
- Antecedents of explainable AI
- Social nature of explainable AI
- Tradeoff between performance and whiteboxing
- Going forward – job opportunities for explainers



Xkcd/1838

16

Difficulty in explainability is nothing new…

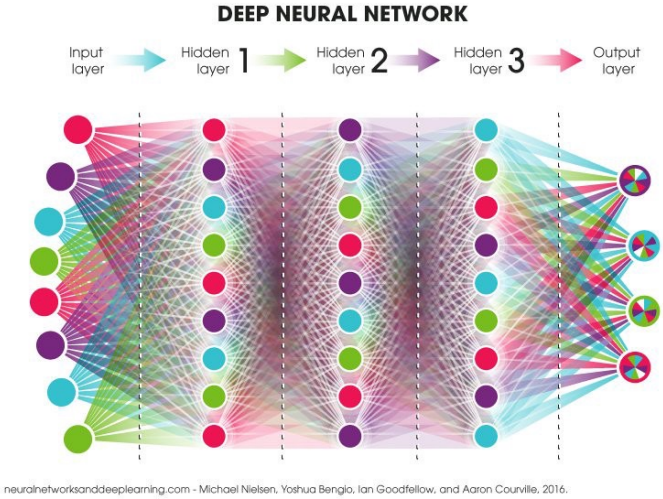
- Air France flight 296 on June 26, 1988 was the first Airbus A320 passenger flight and first public demonstration of a civilian fly-by-wire aircraft
- Mission was to do a low-speed flyover at 30 meters over Mulhouse-Habsheim airport
- Aircraft touched the treetops of the forest at the end of the runway and crashed, killing three passengers
- Even with both flight recorders (digital flight recorder and cockpit voice recorder) at their disposal, due to the complexity of the aircraft automation technology, the accident investigators could not determine whether the aircraft was engaged in stall avoidance mode (i.e., putting the nose of aircraft downwards)

➢ Investigators needed to simulate the flight conditions to examine whether the automatic stall avoidance mode was detected

Photo from: DocumentingReality

17

… however, recent machine learning tools aggravate these difficulties



18

Explainability vs. interpretability

- Interpretability: “being able to ‘translate’ things to understandable form”, often technical, not necessarily outward oriented, and not necessarily stakeholder connected
- Explainability: “bringing things down to certain level”, outward-oriented communication, depending on stakeholders

“Interpretability is about the extent to which a cause and effect can be observed within a system. Or, to put it another way, it is the extent to which you are able to predict what is going to happen, given a change in input or algorithmic parameters. It’s being able to look at an algorithm and go yep, I can see what’s happening here.

Explainability, meanwhile, is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. It’s easy to miss the subtle difference with interpretability, but consider it like this: interpretability is about being able to discern the mechanics without necessarily knowing why. Explainability is being able to quite literally explain what is happening.” https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

19

Reasons for explainable AI

- Human curiosity
  - Humans simply want to understand how things work
- Safety issues
  - Organizations emphasizing reliability might want to refrain from using blackboxed solutions
- Process improvement
  - Blackboxing does not offer fertile ground for process improvement
- Responsible and sustainable AI: Ethical concerns of AI
  - Biases and discrimination associated with blackboxed algorithm development
- Legislative reasons
  - Danish Business Authority needs to explain how companies are identified for fraudulous activitives
  - GDPR issues

20

Responsible and sustainable AI

- Explainability: "In necessary cases, use non-blackboxed models so intermediate steps are interpretable and outcomes are clear, providing transparency to the process."
- Accountability: "Explicit identification of which decisions are delegated to machines, which decisions require human intervention, and who is accountable in either case."
- Fairness: "Must assure AI solutions are balanced and not biased. Need to understand why decisions are made. Need protection against data bias."
- Symmetry: "Must make sure that our data is an asset to us as it is to others."

Daugherty & Wilson (2018). Human+machine. Reimaging work in the age of AI

21

Antecedents of explainable AI

1. Transparency
2. Domain sense
3. Consistency
4. Parsimony
5. Generalizability
6. Trust/performance
7. Fidelity

The following seven slides are modified from: Ahmad, Eckert, Teredesai, Kumar. (2018) Explainable Models for Healthcare AI

22

1. Transparency

- Ability of the machine learning algorithm, model and the features to be understandable by the user
  - The whole model must be understandable simultaneously
    - E.g. A linear model with simple features vs. a linear model with highly engineered features
- Feedback transparency refers to how change in the model will affect the model prediction
- Distinguish between
  - Transparent
    - Regression models
    - Rules-based models
  - Non-transparent
    - Deep learning
    - Gradient boosting models
- LIME (Locally Interpretable Model-Agnostic Explanations)
  - https://homes.cs.washington.edu/~marcotcr/blog/lime/
  - https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

23

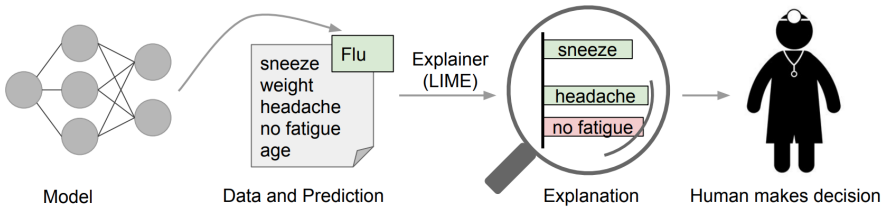


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

24

2. Domain sense

- Explanation should make sense in the domain of application
- Explanations need to be in the right language and also in the right context (Doshi-Velez 2014)
- Making domain sense may require sacrificing or deemphasizing other requirements for explanations such as generalizability
- Interpreting output from machine learning models may also have an element of subjectivity
  - Early Warning Europe example
- Risk assessment often requires domain sense
  - Score 0.5, what does that mean?
- Actionability is often context and role dependent

25

3. Consistency

- Explanation should be consistent across different models and across different runs of the model
- Explanations that are produced by multiple explainable algorithms should be very similar if not the same
- Wide divergence in explanations is a sign of problem with explanations or with the algorithm(s)
- Humans can evaluate quality of explanations across models
  - Scalability issues

26

4. Parsimony

- Explanation should be as simple as possible
  - Applies both to the complexity of the explanation and the number of features provided to explain
  - However, the simplest explanation is not always the best explanation
- Occam's razor…
  - … demands that scientists accept the simplest possible theoretical explanation for existing data
  - Razor refers to "shaving away" unnecessary assumptions or cutting apart two similar conclusions

27

5. Generalizability

- A good algorithm is generalizable
- Distinguish between
  - Local models
  - Cohort (a group of people who share a common characteristic over a certain period of time) level models
  - Global models
    - Decision trees, rule-based models etc.
- Danish business authority is collaborating with European Commission and taking the Early warning algorithm to Poland, Greece, Italy and Spain

28

6. Trust/performance

- Expectation that the corresponding predictive algorithm for explanations should have a certain performance
- Explanations accompanied with sub-par predictions can foster distrust
- Tradeoff between interpretability and accuracy (discussed later)

29

7. Fidelity

- Expectation that explanation and predictive model align well with one another
- Explanation will be as good as the data
- Incorrect explanations may result from problems in the data
- Constraints on data collected may also show up as constraints in explanations
- Explanation is *sound* if it adheres to how the model actually works
- Explanation is *complete* if it encompasses the complete extent of the model

30

Social nature of explainable AI (explainer vs. explainee)

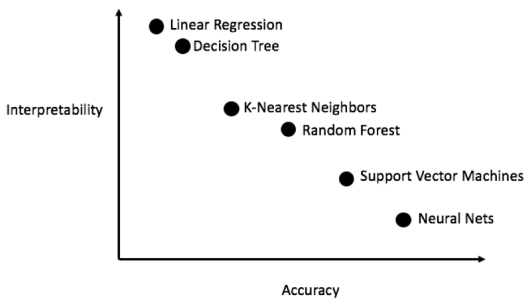
- Three layers of social explanation (Malle 2004)
  - Conceptual framework that outlines the *assumptions* people make about human behaviour and explanation
  - *Psychological processes* that are used to construct explanations
  - *Language layer* that specifies the type of linguistic structures people use in giving explanations
  - ➢ Creating a shared meaning is important for explaination of AI
- Case Danish business authority
  - Need to explain primarily emerges from the requirement to explain the work of the algorithm among human experts (e.g. developer-lawyer; lawyer-client organization)

B. F. Malle, How the mind explains behavior: Folk explanations, meaning, and social interaction, MIT Press, 2004.

31

Tradeoff between accuracy and whiteboxing

- Interpretability of neural nets often lower than for linear regression
  - Trade-off between accuracy and interpretability
- Tradeoff often between accuracy, explanation and risk
  - In high-risk domains, need of explanation is high
  - In low-risk domains, reduced need to explain so optimization centered on prediction accuracy
- Performance, however, often necessitates high usability which, in turn, requires interpretability



https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9

32

Jobs for explainers

- Future job opportunities will open up to bridge the gap between technologists and business leaders and these jobs will become more important as AI systems become more opaque and blackboxed.
- Managers seek explanations especially in those circumstances where systems recommend action that may go against the grain of conventional wisdom or that could be controversial.

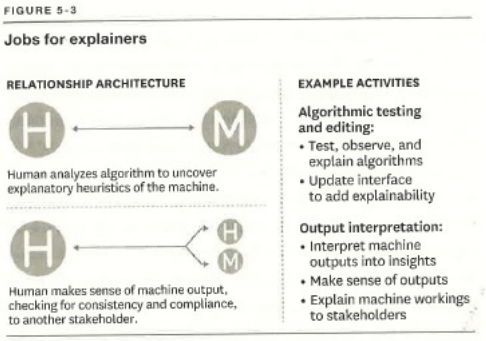


FIGURE 5-3

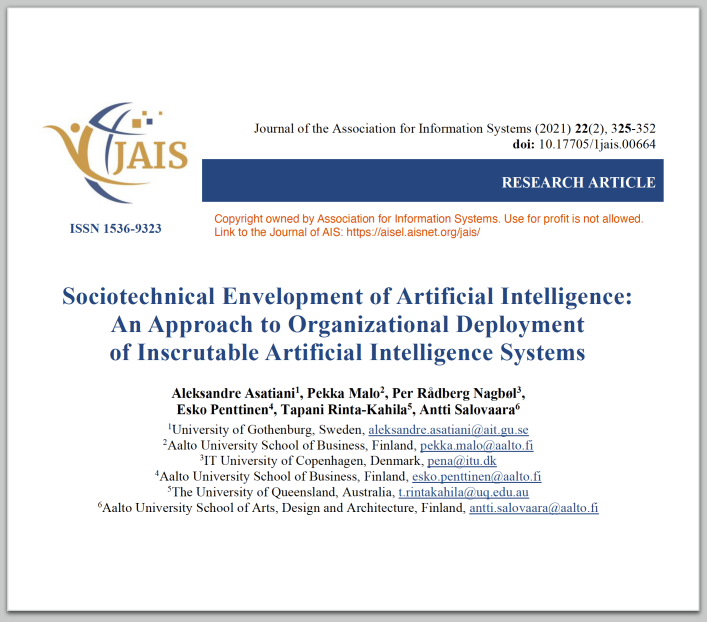
Jobs for explainers

RELATIONSHIP ARCHITECTURE	EXAMPLE ACTIVITIES
H ⟷ M Human analyzes algorithm to uncover explanatory heuristics of the machine.	**Algorithmic testing and editing:** • Test, observe, and explain algorithms • Update interface to add explainability
H ⟶ H, M Human makes sense of machine output, checking for consistency and compliance, to another stakeholder.	**Output interpretation:** • Interpret machine outputs into insights • Make sense of outputs • Explain machine workings to stakeholders

Daugherty & Wilson (2018). Human+machine. Reimaging work in the age of AI

33

How to deploy AI safely – Case DBA



JAIS
ISSN 1536-9323
Journal of the Association for Information Systems (2021) 22(2), 325-352
doi: 10.17705/1jais.00664
RESEARCH ARTICLE

Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems

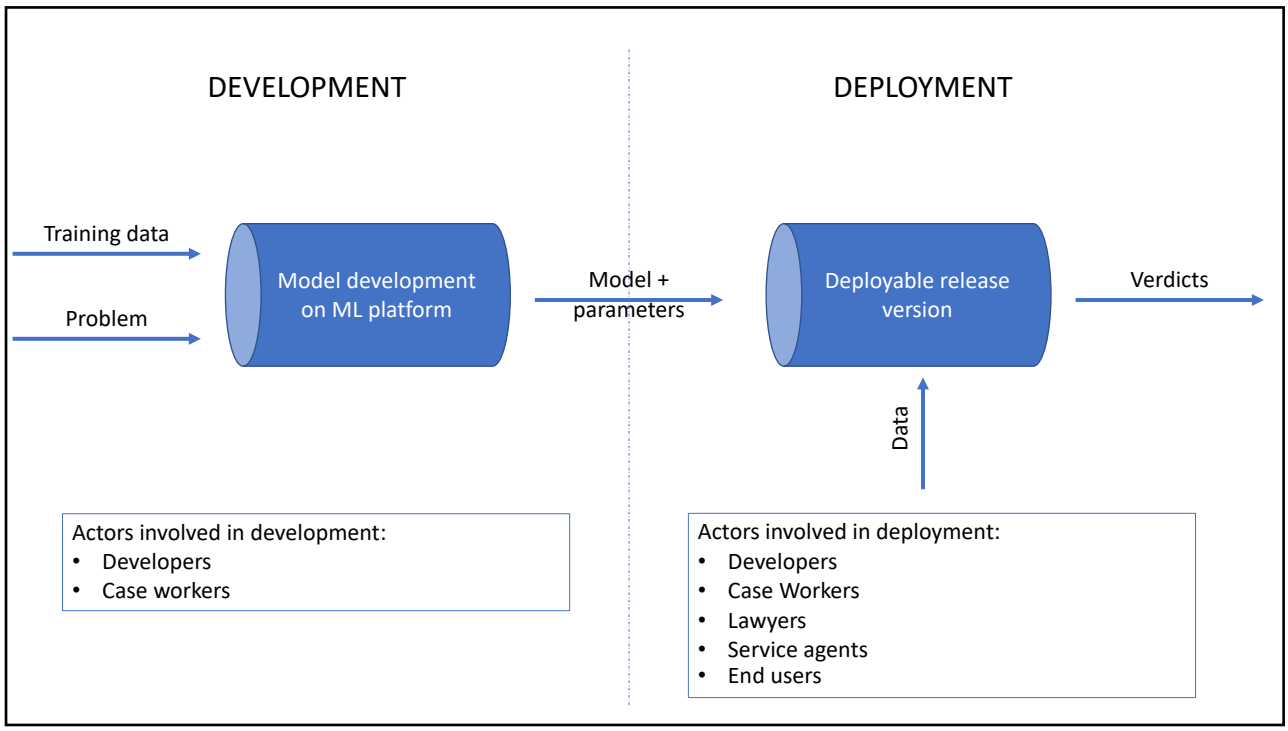
Aleksandre Asatiani[1], Pekka Malo[2], Per Rådberg Nagbøl[3], Esko Penttinen[4], Tapani Rinta-Kahila[5], Antti Salovaara[6]
[1]University of Gothenburg, Sweden, aleksandre.asatiani@ait.gu.se
[2]Aalto University School of Business, Finland, pekka.malo@aalto.fi
[3]IT University of Copenhagen, Denmark, pena@itu.dk
[4]Aalto University School of Business, Finland, esko.penttinen@aalto.fi
[5]The University of Queensland, Australia, t.rintakahila@uq.edu.au
[6]Aalto University School of Arts, Design and Architecture, Finland, antti.salovaara@aalto.fi

34

Machine learning at the Danish Business Authority (DBA)

- Danish Business Authority (DBA) is a Danish government unit with regulatory obligations related to supervision of Danish companies and fraud prevention
- Extensive use of structured data (as opposed to paper and PDF) in financial statements (XBRL) has paved the way for data analytics at DBA
- Numerous machine-learning projects on-going at DBA
  - Many of these projects include the use of intractable systems; however, as public organization, DBA must be able to explain how their decisions are made
  - How can DBA use these systems without explainability issues spiralling out of control?

35



36

ML projects at the DBA

Project name	Project description [use case within DBA, end users]	Purpose	Input	Output	Model and tool
Auditor's Statement	The Auditor's Statement model speeds up verification that the valuations of company assets given in an auditor's statement are correct and that the statement does not feature violations. The algorithm is used by internal DBA case workers.	Prevent misreporting of company assets	Text from auditor's statements that present asset valuations	Probability of violations in asset valuations	Random forest, bag of words
Bankruptcy	The Bankruptcy model predicts company distress and insolvency and ties in with the Early Warning Europe (EWE) initiative. The algorithm is used not at the DBA but by external consultants in the EWE community in Denmark and in the European Union. The DBA is not responsible for actions and consequences related to the tool.	Identify companies in distress to enable timely intervention	Data from the business registry and annual financial reports	Probability of bankruptcy	Scikit-learn, gradient boosting
Company Registration	The Company Registration model is aimed at detecting fraud-indicating behavior among newly registered Danish companies. The algorithm is used by internal DBA case workers.	Prevent abusing incorporation to commit fraud	Data from the business registry, annual reports, and VAT reports	Probability of fraudulent actions	XGBoost
Land and Buildings	The Land and Buildings model predicts violations of accounting policies related to property holdings and long-term investments. The algorithm is used by internal DBA domain experts.	Prevent violations of accounting policy	Text about accounting policies, from the auditor's statement	Probability of violations of accounting policies	Random forest, bag of words
Passport	The Passport model expedites the processing of submitted documents by supplying a text string from the machine-readable portion of a passport and comparing it against input data from the user. The algorithm is used by internal DBA case workers.	Facilitate processing of documents	Pictures of IDs submitted to the DBA	JSON string with text from the machine-readable portion of the ID	PassportEye
Recommendation	The Recommendation model improves the user experience of the DBA's virk.dk online portal by focusing on personalized content and optimized interfaces. The algorithm improves the portal's usability for external customers (end users).	Improve usability of the online portal	Telemetry data from virk.dk	Recommendation of relevant content	TBD
Sector Code	The Sector Code model speeds up verifying a company's industry-sector code. At present, 25% of the company codes are incorrect. The algorithm is used by internal DBA case workers.	Prevent misreporting of industry sector codes	Activity-description text from a company's annual statements	Probability distribution over the set of sector codes	Neural network
Signature	The Signature model, in combination with the associated document filter, speeds up verification of whether a company founding document is signed or not. The algorithm is used by internal DBA case workers and returns three probabilities: of whether the document is physically signed, whether it is digitally signed, and whether the signature is missing.	Facilitate the process of founding a company	An image of a company-establishment document	Probability of whether a document is signed or not	Neural network (ResNet16)

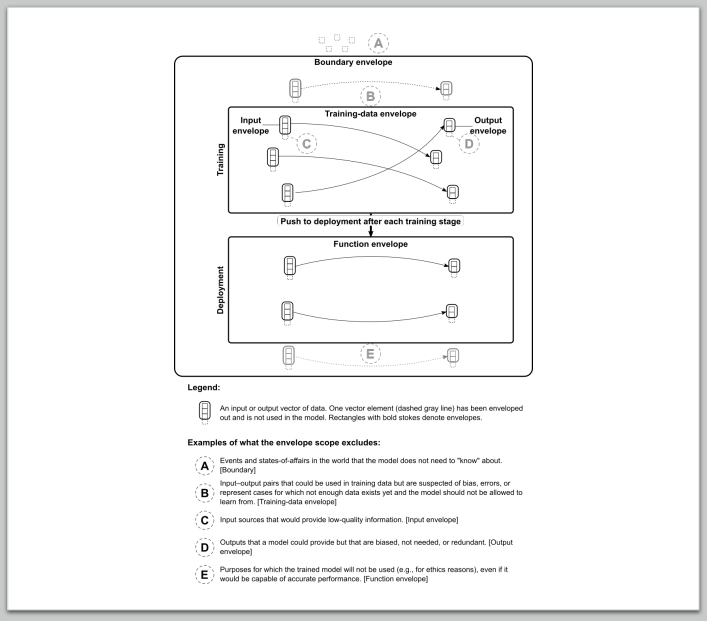
37

One proposed solution is envelopment

- Recently, the notion of envelopment has been proposed as a tool to improve our understanding on responsible regulation and use of AI (Robbins 2019)
  - Envelopment concept originates from early research on physical robotics where it referred to the "the set of points representing the maximum extent or reach of the robot hand or working tool in all directions" expressed as shaded regions in factories' floor maps
  - In AI and algorithmic work, envelopment may manifest through control on training data, setting boundaries on algorithm and controlling its inputs, and knowing the functions and outputs of AI (Robbins 2019)
- Case study on DBA
  - Envelopment methods: input & output data control, training data control, control on boundaries, function control

38

Envelopment



39