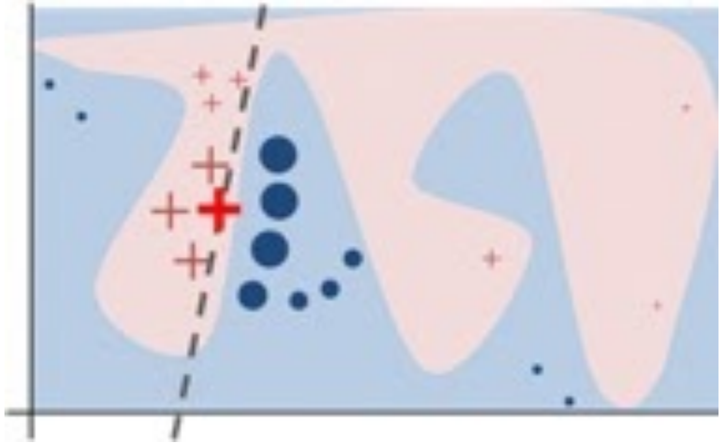
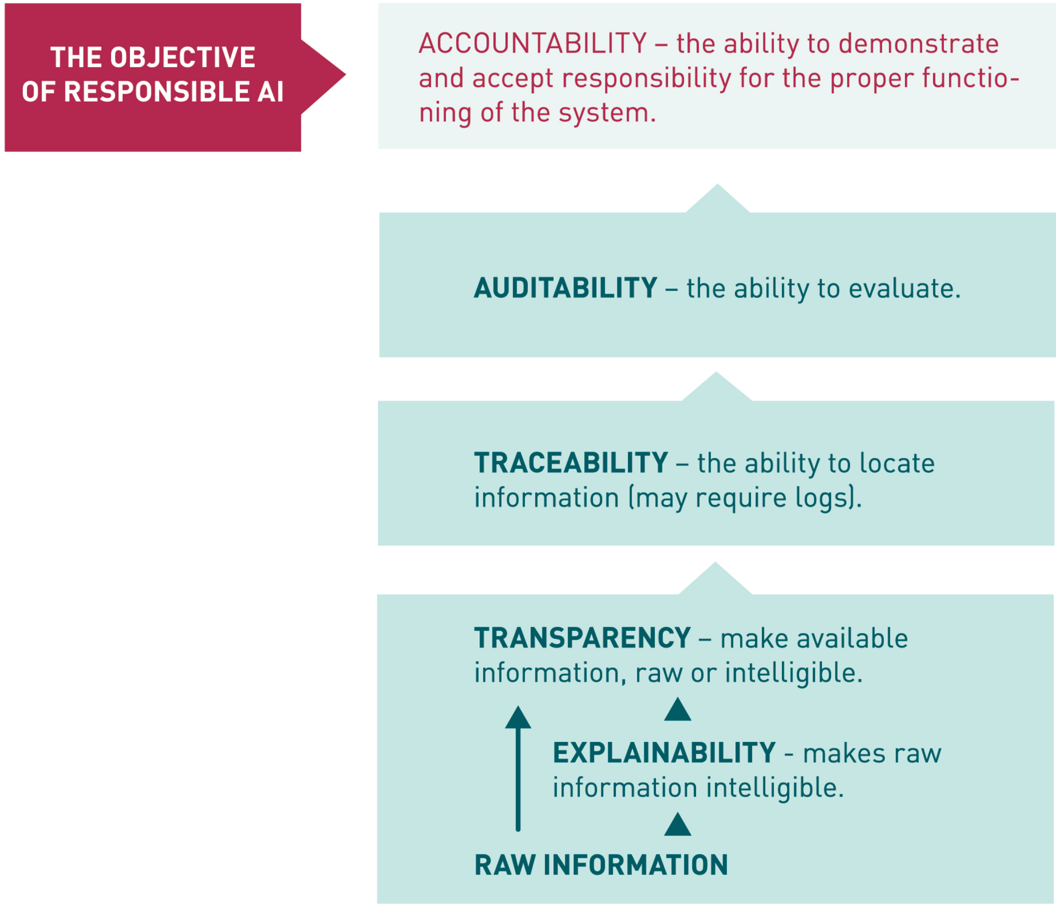
Explainability

Winston Maxwell

1er décembre 2020



FROM TRANSPARENCY TO ACCOUNTABILITY



THE OBJECTIVE OF RESPONSIBLE AI

ACCOUNTABILITY – the ability to demonstrate and accept responsibility for the proper functioning of the system.

AUDITABILITY – the ability to evaluate.

TRACEABILITY – the ability to locate information (may require logs).

TRANSPARENCY – make available information, raw or intelligible.

EXPLAINABILITY - makes raw information intelligible.

RAW INFORMATION

THE THREE PILLARS OF EXPLAINABILITY

CONTEXT FACTORS

RECIPIENT

- Who is receiving the explanation,
- Level of expertise,
- Time available.

IMPACT

- What harms possible,
- Can explanation mitigate harm.

REGULATORY

- What regulatory framework,
- Fundamental rights.

OPERATIONAL

- Is explainability an operational imperative,
- Safety certification,
- Usability need.

TECHNICAL SOLUTIONS

POST-HOC

- LIME,
- Kernel-SHAP,
- Saliency maps.

HYBRID

- Modifying objective or predictor function,
- Produce fuzzy rules,
- Output approaches,
- Input approaches,
- Genetic fuzzy logic.

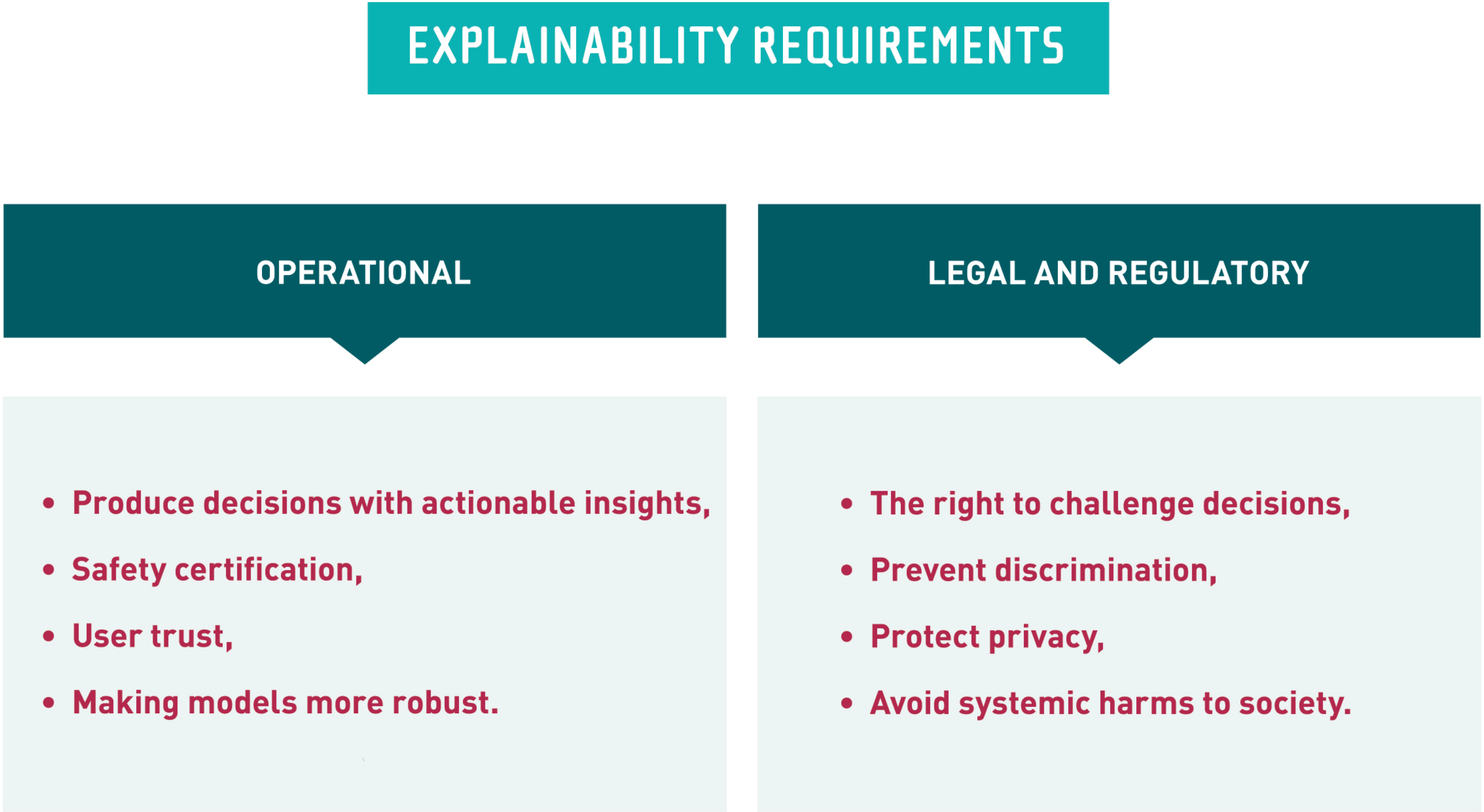
EXPLAINABILITY PARAMETERS

GLOBAL EXPLAINABILITY

- User’s manual,
- Level of detail,
- Source code,
- Info on training data,
- Learning algorithm,
- Disclosure of biases,
- Copy of training data.

LOCAL EXPLAINABILITY

- Counterfactual dashboards,
- Saliency maps,
- Level of detail,
- Individual decision logs,
- What information in logs, and store how long.



EXPLAINABILITY REQUIREMENTS

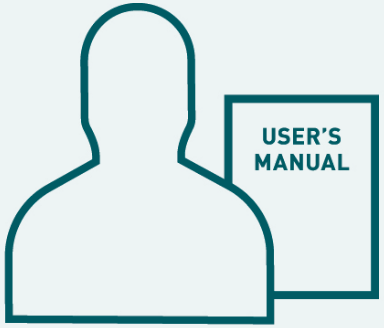
OPERATIONAL

- Produce decisions with actionable insights,
- Safety certification,
- User trust,
- Making models more robust.

LEGAL AND REGULATORY

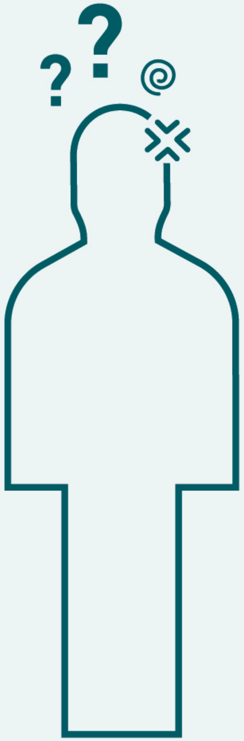
- The right to challenge decisions,
- Prevent discrimination,
- Protect privacy,
- Avoid systemic harms to society.

GLOBAL EXPLANATIONS



- Overview of the model,
- How it learned,
- What training data,
- Limitations to the model and use restrictions.

LOCAL EXPLANATIONS



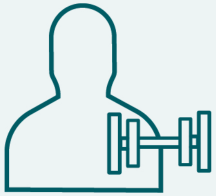
- Why was my application denied?
- What was the main factor?
- What can I change?
- Was I discriminated against?
- Why did you misclassify this image?

EXPLAINABILITY REQUIRES INFORMATION ABOUT THE LEARNING ALGORITHM AND THE TRAINED ALGORITHM

QUESTIONS TO ASK THE LEARNING ALGORITHM

- How did you learn?
- What's your objective function?
- What training and test data did you use?
- How did you tune the model?

QUESTIONS TO ASK THE TRAINED ALGORITHM

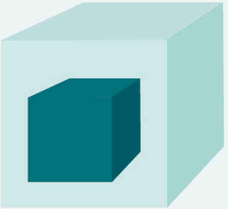


- What were the main factors leading to your decision?
- How would the decision change if we altered one of the factors?
- Show me a map of how you reasoned in this case.

TECHNICAL APPROACHES

POST-HOC APPROACHES

- Input perturbation,
- Saliency maps.



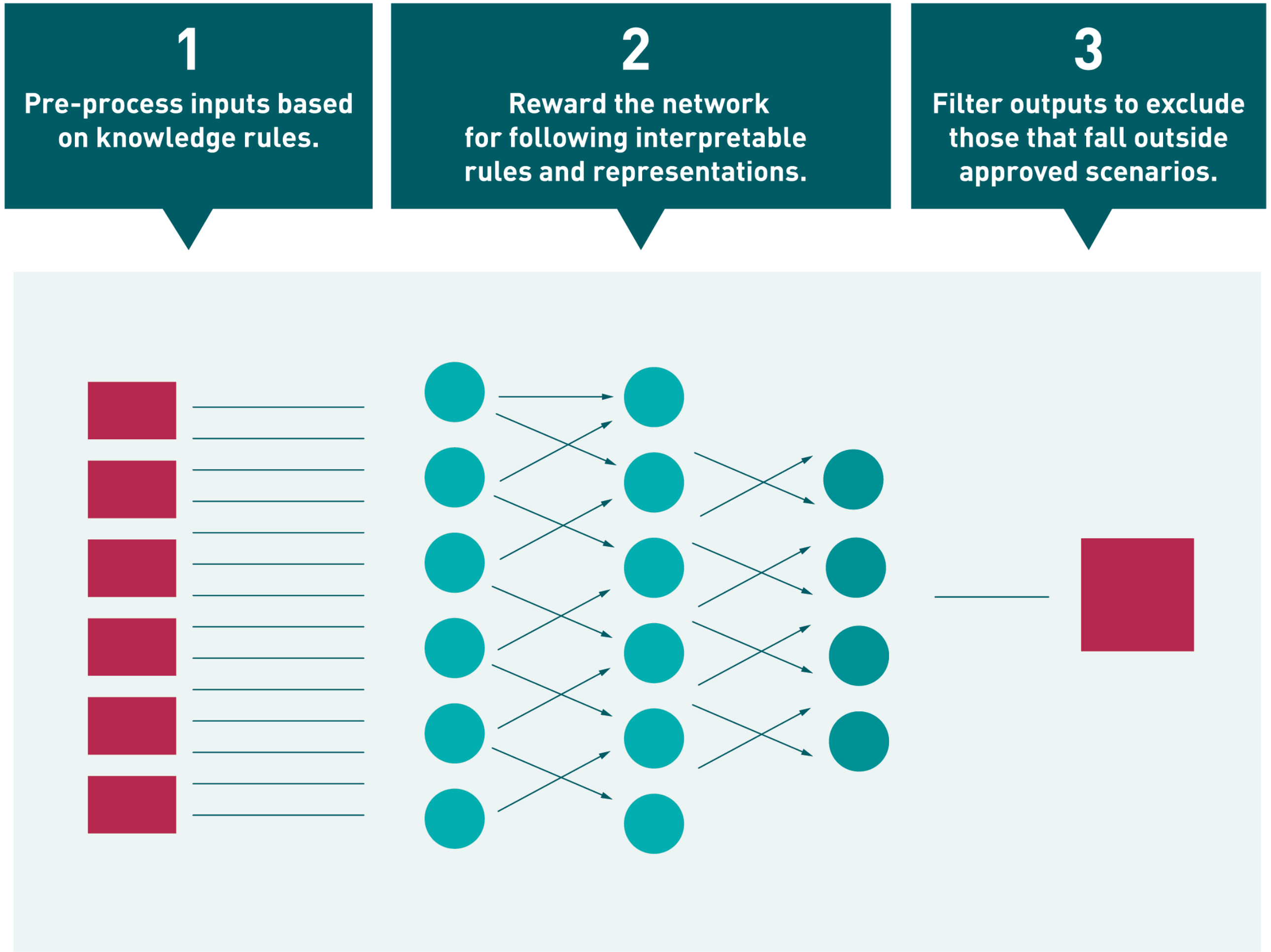
Explanor model produces local explanations by approximating the black box function.

Black box model remains black, and continues to make predictions without explanation.

PUTTING EXPLAINABILITY IN THE MODEL

Modifying the black box itself to make it more understandable.

THREE APPROACHES TO EXPLAINABILITY BY DESIGN



1

Pre-process inputs based on knowledge rules.

2

Reward the network for following interpretable rules and representations.

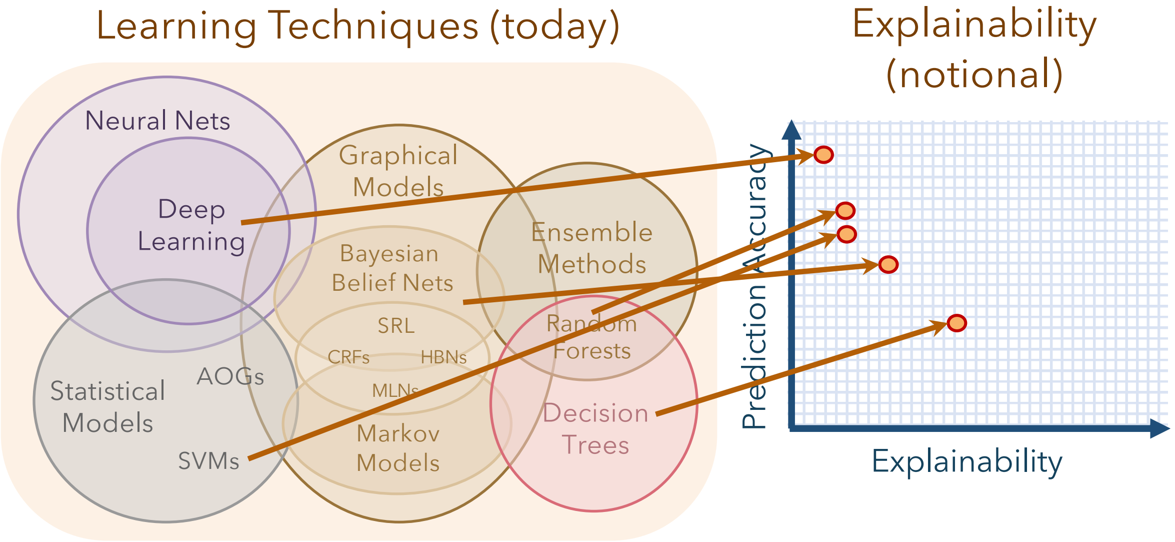
3

Filter outputs to exclude those that fall outside approved scenarios.

B.1 Explainable Models

New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance



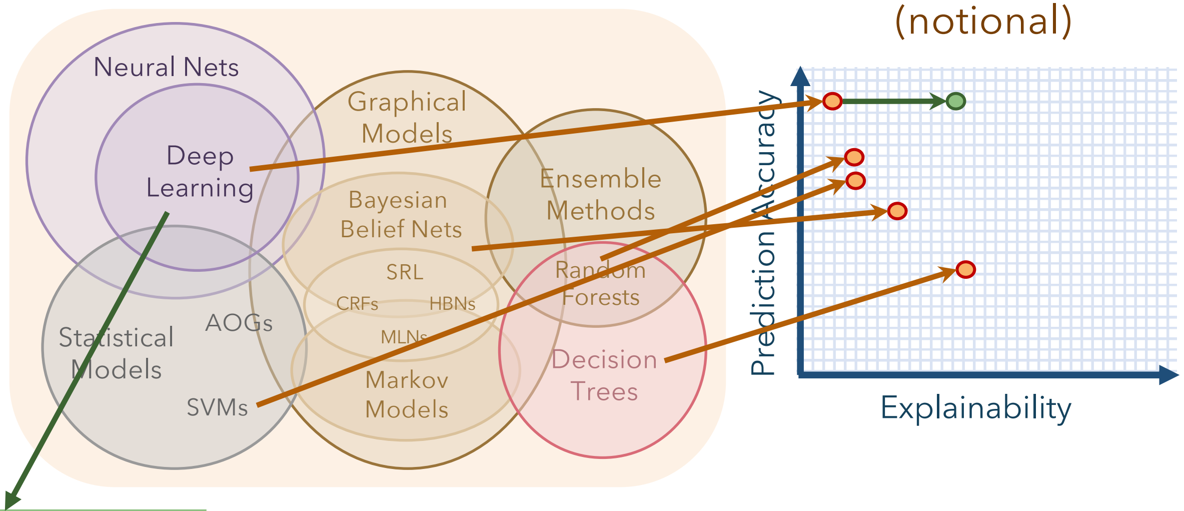
B.1 Explainable Models

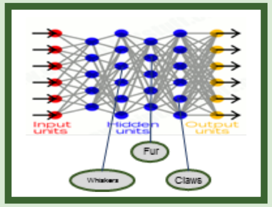
New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)

Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features

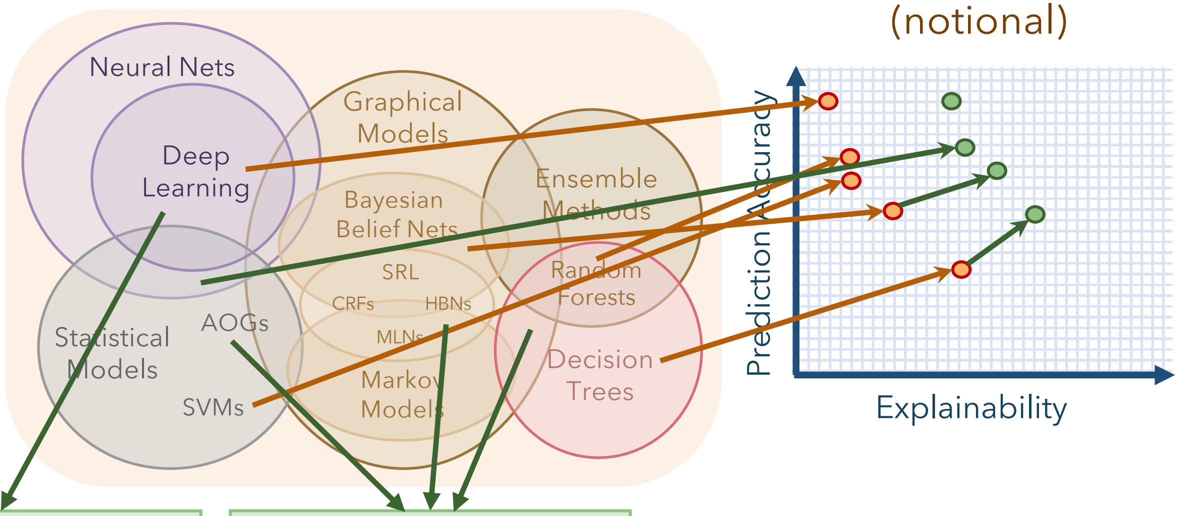
B.1 Explainable Models

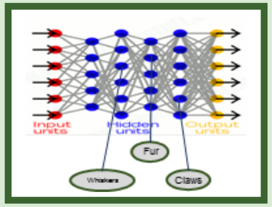
New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Learning Techniques (today)

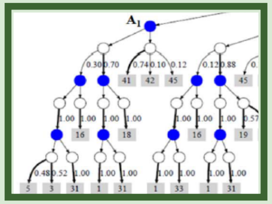
Explainability (notional)





Deep Explanation

Modified deep learning techniques to learn explainable features



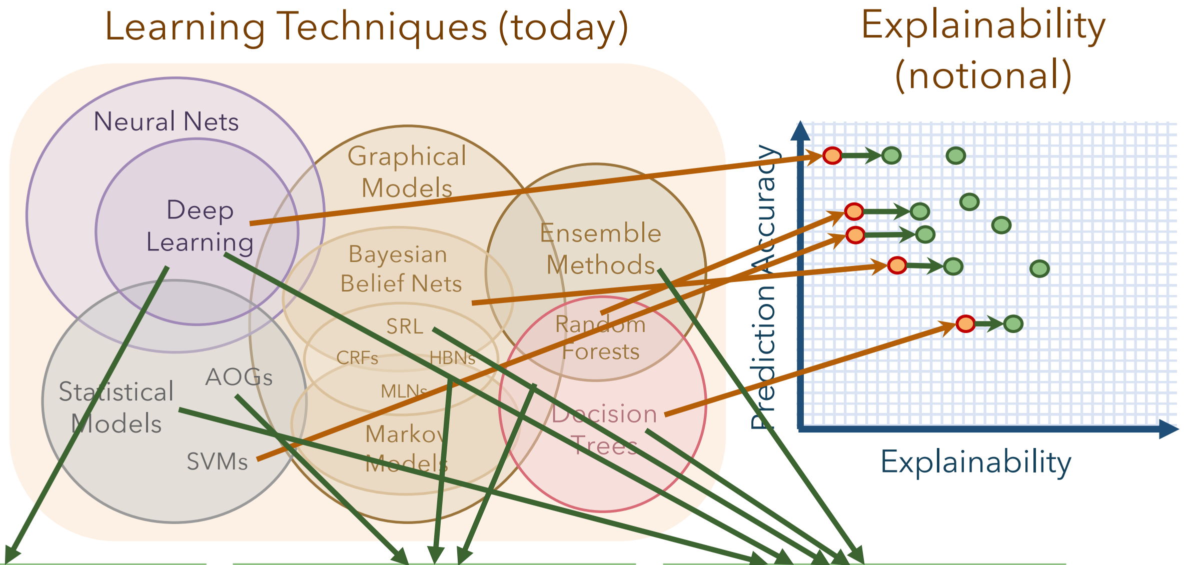
Interpretable Models

Techniques to learn more structured, interpretable, causal models

B.1 Explainable Models

New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance



Learning Techniques (today)

Neural Nets

Deep Learning

Graphical Models

Bayesian Belief Nets

Ensemble Methods

Random Forests

SRL

CRFs

HBNs

MLNs

Markov Models

Statistical Models

AOGs

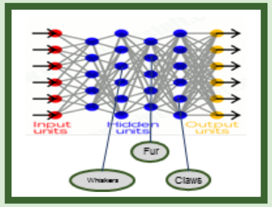
SVMs

Decision Trees

Explainability (notional)

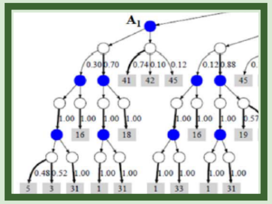
Prediction Accuracy

Explainability



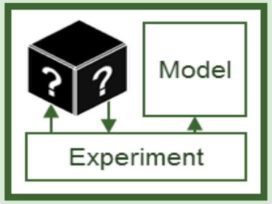
Deep Explanation

Modified deep learning techniques to learn explainable features



Interpretable Models

Techniques to learn more structured, interpretable, causal models



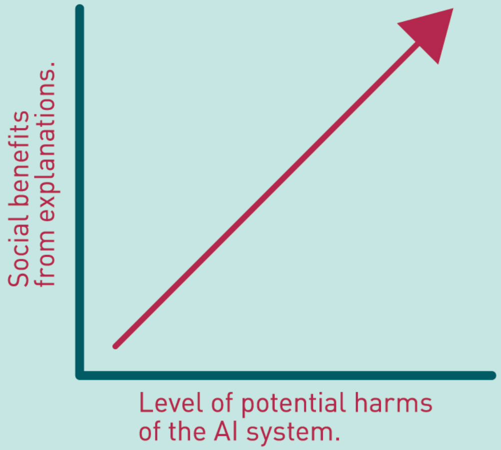
Model Induction

Techniques to infer an explainable model from any model as a black box

BENEFITS AND COSTS OF EXPLANATIONS

BENEFITS

Benefits of explanation generally increase with the level of potential harm caused by AI.



COSTS

- Design and operations costs,
- Reverting to less-performing models,
- Creating and storing logs,
- Interference with trade secrets,
- Interference with other rights, e.g. data protection.

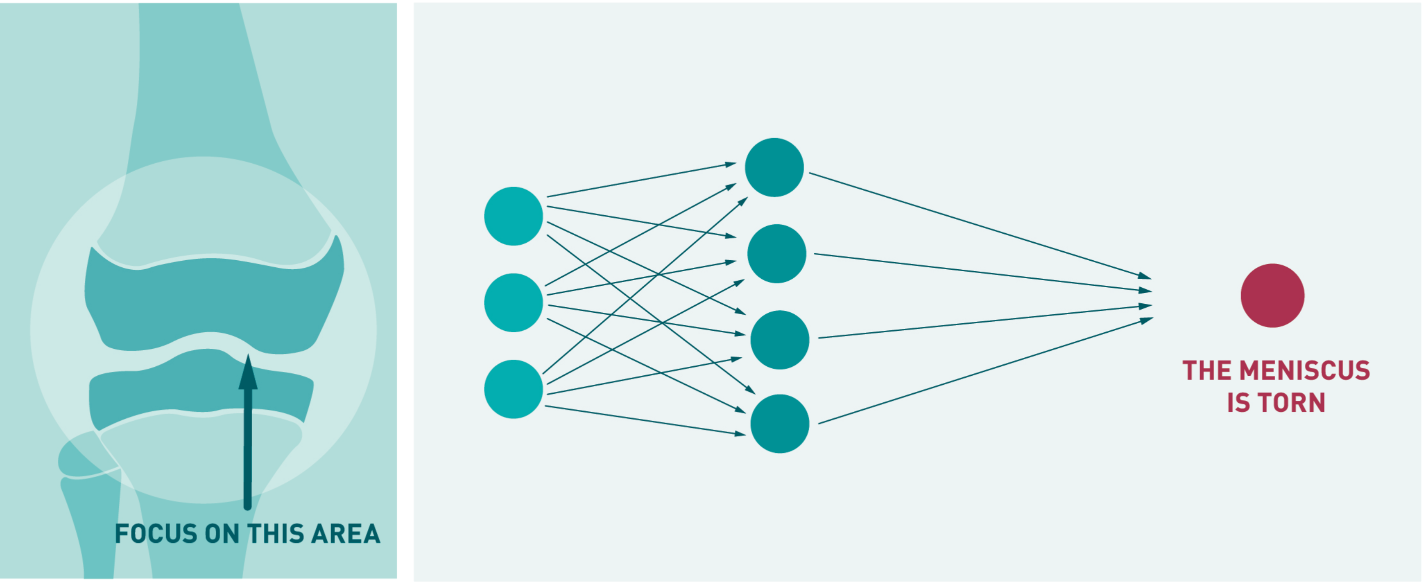
EXAMPLE OF A HYBRID APPROACH

Knowledge rules tell the network what part of the image to focus on when learning.

X-RAY OF A KNEE

NEURAL NETWORK

PREDICTION



FOCUS ON THIS AREA

THE MENISCUS IS TORN

Explainability and the law

A fundamental right?
Charter of fundamental rights
Human dignity – art 1
Privacy/data protection– arts 7 et 8
Non-discrimination – arts 20, 21 et 23
Effective remedy – art. 47
=>Déc. Conseil Constitutionnel 2018-765 DC du 12 juin 2018, para. 65 à 72

- **GDPR (articles 22 et 12, 13 et 14)**
- **La loi Lemaire**
  - **Public administrations (Article R311-3-1-1 et 2 du code des relations entre le public et l'administration)**
  - **Plateformes (D 111-7 code de la consommation)**
- **« Platform to business » Regulation (Regulation EU 2019-1150, articles 5 et 7, recitals 24 à 28)**

Question: Does explainability require communication of source code?

Explainability and the law

- **The CJEU « Quadrature du Net » case**

*« Furthermore, since the automated analyses of traffic and location data necessarily involve some margin of error, any positive result obtained following automated processing must be subject to an **individual re-examination by non-automated means** before an individual measure adversely affecting the persons concerned is adopted, such as the subsequent real-time collection of traffic and location data, since such a measure cannot be based solely and decisively on the result of automated processing. Similarly, in order to ensure that, in practice, the pre-established models and criteria, the use that is made of them and the databases used are not discriminatory and are limited to that which is strictly necessary in the light of the objective of preventing terrorist activities that constitute a serious threat to national security, a regular re-examination should be undertaken to ensure that those pre-established models and criteria and the databases used are reliable and up to date (see, to that effect, Opinion 1/15 (EU-Canada PNR Agreement) of 26 July 2017, EU:C:2017:592, paragraphs 173 and 174). »*

Explainability and the law

- ***NJCM v. the Netherlands*, District Court of The Hague, Case n° C-09-550982-HA ZA 18-388, February 5, 2020.**

- **Netherlands law authorized use of algorithm to predict score of social security fraud.**
- **Algorithm is fed by several separate government data bases.**
- **Law provided for safeguards, including requirement of human intervention, and institutional supervision.**
- **Court found violation of ECHR.**
- **WHY?**
  - **Profiling is serious interference with privacy right**
  - **Therefore needs to pass proportionality test**
  - **One element of proportionality test is to make sure that there are sufficient safeguards to assure « fair balance »**
  - **Transparency of the algorithm is an important safeguard**
  - **Lack of transparency prevents individuals from challenging their score and prevents courts and regulators from verifying absence of discrimination**

Explainability and the law

- ***State of Wisconsin v. Loomis*, Supreme Court of Wisconsin, 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S.Ct. 2290 (2017)**
- COMPAS calculates a probability that a particular person will commit another crime if released.
- The system was used by a judge as a source of information to fix the sentence itself. The affected individual, Mr. Loomis, argued that the COMPAS system is opaque, that the source code is unavailable, and that the system is racially biased.
- The Supreme Court of Wisconsin found that the judge's use of the algorithm **did not affect Mr. Loomis's constitutional right to due process** because the algorithmic score was an insignificant element in the judge's decision, the judge relying almost exclusively on other factual elements.
- But the court said that algorithmic tools like COMPAS must be accompanied by a warning statement on the algorithm's limitation, including disclosure on the population sample used to train the scoring system and the fact that the population sample may not correspond to the relevant local population. The documentation should disclose that the owners of the algorithm refuse to give access to source code, that studies have shown the system to disproportionately classify minority offenders as having a higher risk of recidivism, and that the system needs to be "**renormed for accuracy due to changing populations and subpopulations**."
- The court emphasized that an algorithm may in no case be used to determine the sentence itself, but only as a source of information on how the sentence should be served.

Explainability and the law

- ***Local 2415 v. Houston Independent School District*, 251 F. Supp. 3d 1168 (S.D. Tex. 2017).**

- An algorithmic scoring system used to rank teachers had to be open to scrutiny in order to permit the affected teachers to verify the accuracy of their score and challenge the decision if they disagree.
- The court found that without access to “value-added equations, computer source codes, decision rules, and assumptions”, teachers could not exercise their constitutionally-protected rights to due process.
- The court said that teachers needed to be able to replicate the algorithmic decision to verify whether there was an error in the score. Without this ability, the teachers’ scores remain “**a mysterious ‘black box’, impervious to challenge**.”

https://www.co ments/2020/10-octobre/sre_dauphine_dossier-participant_version-finale

- **Washington State Facial Recognition Law**

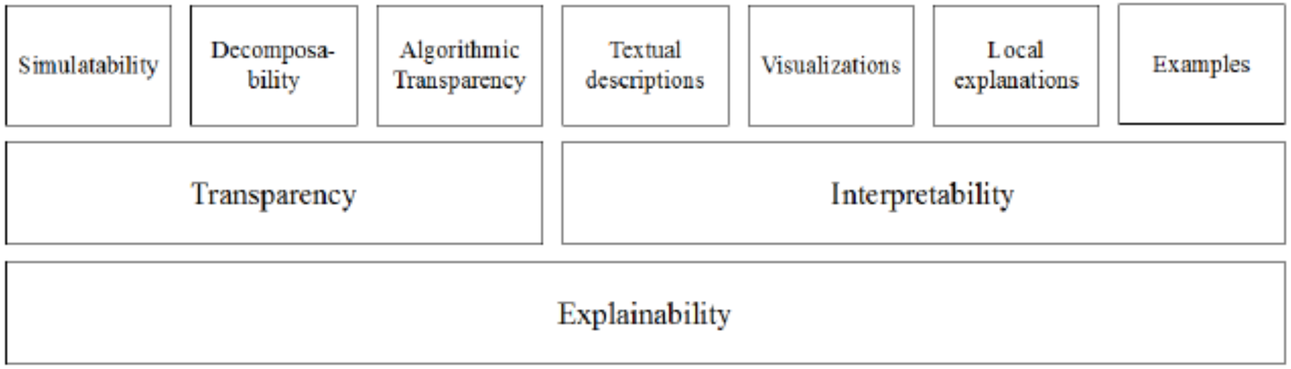
- The name of the facial recognition service, vendor, and version; and (ii) a description of its general capabilities and limitations, including reasonably foreseeable capabilities outside the scope of the proposed use of the agency**;**
- The type or types of data inputs that the technology uses; (ii) how that data is generated, collected, and processed; and (iii) the type or types of data the system is reasonably likely to generate;
- A description of the purpose and proposed use of the facial recognition service, including what decision or decisions will be used to make or support it; (ii) whether it is a final or support decision system; and (iii) its intended benefits, including any data or research demonstrating those benefits;
- **Information on the facial recognition service's rate of false matches, potential impacts on protected subpopulations, and how the agency will address error rates, determined independently, greater than one percent;**
- **A description of any potential impacts of the facial recognition service on civil**

Explainability and human decisions

- **How to ensure a « meaningful human decision »?**
  - GDPR recital 71
  - *In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.*
- **The problem(s) of automation bias**
  - **Over-reliance on AI recommendations**
    - Face matching
    - GPS
  - **Under-reliance**
    - Rejection of decision tool (disuse)
- **Adversarial explanations?**

Explainability has different angles

- Transparency means access to the blueprint and inner workings of the algorithm
  - Source code, training data, test data, data cleaning, tuning of the algorithm,
  - But this information will remain meaningless in many situations, particularly for non-specialists
- Interpretability means providing explanations that are meaningful for humans, and that can answer important questions like:
  - Why did the classifier make this particular decision? (local explainability)
  - What factors were the most important? What were the weights given to those factors?
  - What if we changed one of those factors? (counterfactual explainability)
  - What is the logic of the system as a whole? (global explainability)



Simulatability | Decomposability | Algorithmic Transparency | Textual descriptions | Visualizations | Local explanations | Examples

Transparency | Interpretability

Explainability

Figure 2: Taxonomy of explainability in the field of ADM.

Source: Waltl & Vogel, Explainable Artificial Intelligence, 2018

Explainability in the law

- Fair Credit Reporting Act (FCRA), Equal Credit Opportunity Act (ECOA), Regulation B require ‘adverse action notices’:
  - If a creditor takes an adverse against an applicant, the creditor must give a statement of ‘specific reasons’ for the denial. Same obligation for employment decisions based on credit information.
  - Regulation B provides a list of 24 reason codes
- Mortgage Credit Directive (art. 18) and Consumer Credit Directive (art. 9) require tranparency in case of rejection based on automated processing of data or the consultation of a database

Explainability in the law

- **Fundamental rights: explanation may be required under principles of due process, right to judicial review, right to defense, right to data protection.**
  - **Houston Fed'n of Teachers, Local 2415 v. Houston Indep. Sch. Dist., 251 F. Supp. 3d 1168 (S.D. Tex. 2017**
  - ***Alaska v. Lubchenco*, 825 F. Supp. 2d 209 (D.D.C. 2011) at 221.**
  - **Loomis case (Wisconsin)**
  - **CJEU Quadrature du Net Cases C-511/18, C-512/18, C-520/18**
  - **Netherlands Social Security Fraud (SyRI) case**
- **Administrative Procedure Act: administrative decisions ”shall include a statement of . . . findings and conclusions, and the reasons or basis therefor, on all the material issues of fact, law, or discretion presented on the record.” (5 U.S.C. §557(c))**

Explainability in the law

- GDPR requires
  - that a data controller give ‘meaningful information about the logic involved’
  - ‘fair and transparent’ processing
- Accountability / reversal of burden of proof: explanation may be required to escape liability or reduce sanctions: where an entity must affirmatively prove that its system was safe, compliant, non-discriminatory, state-of-the-art, effective controls in place, etc.
  - Reversal of the burden of proof happens when the law puts the burden on the company to affirmatively prove it had effective/safe measures in place
  - the controller ’shall be able to demonstrate that processing is performed in accordance with this Regulation’ (GDPR art. 24)
  - adopt internal policies to implement data protection by design and by default (GDPR art. 25)
  - undertake data protection impact assessments (GDPR art. 35)

Reversal of burden of proof; proof of compliance

- Product liability directive
  - ‘The producer shall not be liable as a result of this Directive if he proves …. (e) that the state of scientific and technical knowledge at the time when he put the product into circulation was not such as to enable the existence of the defect to be discovered’ (Product Liability Directive 85/374/EEC, art. 7(e))
- Res ipsa loquitur doctrine in U.S. torts:
  - presumption of negligence, requiring affirmative proof by the defendant that he or she was not negligent
- U.S. Sentencing Guidelines: ‘effective compliance and ethics program’
  - ‘The organization's governing authority shall be knowledgeable about the content and operation of the compliance and ethics program and shall exercise reasonable oversight with respect to the implementation and effectiveness of the compliance and ethics program.’ (§8.B21)

Legal ecosystem for explainable AI

Washington Facial Recognition Act

FCRA, ECOA

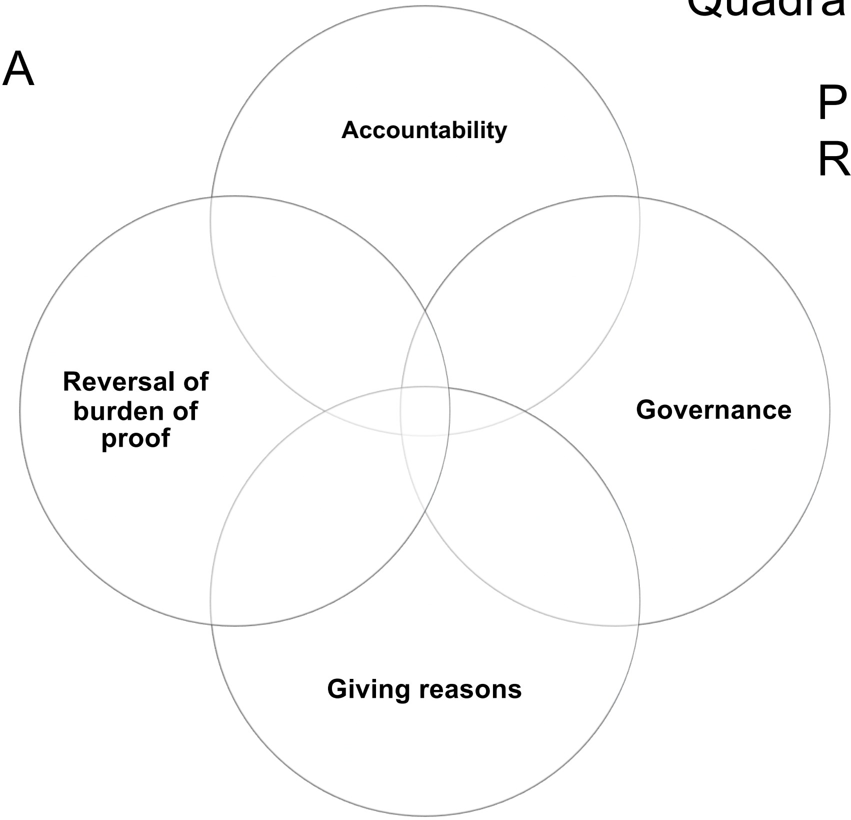
14th Amendment US Constitution (due process)

Administrative Procedures Act

Civil Rights Act

US Sentencing Guidelines

Res ipsa loquitur



Fundamental Rights (eg Quadrature du Net)

Platform to Business Regulation

GDPR

Product Liability Directive

Mortgage Credit Directive, Consumer Credit Directive

As a lawyer, what explainability do you need?

- **Lawyers will be on the front line to show regulators and third parties that the AI system was surrounded by appropriate controls and safeguards to ensure that its operation is safe, compliant, fair, transparent and non-discriminatory.**
- **This involves two tasks:**
  - Creating the appropriate controls and safeguards in the first place;
    - This involves adding a layer of controls and safeguards in order to transform an algorithm that is accurate and replicable for solving a particular mathematical problem into a decision process that is safe, compliant and fair.
  - Documenting the controls and safeguards in a way that is meaningful for regulators/third parties.
- **Before helping on these tasks, a lawyer needs to know enough about AI to ask the right questions, not be confused by data science jargon, and identify weaknesses in algorithmic decision-making.**

Due diligence and documenting policies

- **Recommended steps (source: IEEE Ethically Aligned Designed – interim draft Dec. 2017)**
  - AI systems should have an 'ID Tag'
  - Required documentation on permitted uses, required training
  - Required maintenance
  - See e.g. Washington State Facial Recognition Law
- **Guidelines on outsourcing by the Committee of European Banking Supervisors**
  - Security of data and systems
  - Location of data and data processing
  - Access and audit rights
  - Chain outsourcing
  - Contingency plans and exit strategies