Machine Learning in the banking landscape

Market experience, a view on specific risks, controls and audit considerations under the EU HO2020 program

February 2020

EY 160 JAHRE in Deutschland

EY Building a better working world

Our suggestion for this session

1. Definitions and a bird's eye biew on today's topic
2. ML/AI State of Play as we perceive it in Banks today
3. Where do we see bigger challenges for banks with ML/AI algorithms/models than with traditional ones
   - 3.1 Pioneering somewhat new techno-logy in a shaping environment
   - 3.2 Zooming in on model explainability
   - 3.3 A closer look at the bias topic – fundamental concepts and application
   - 3.4 Conclusions, if any?
4. Selected aspects: What we would look for when inspecting AI/ML algorithms in financial institutions

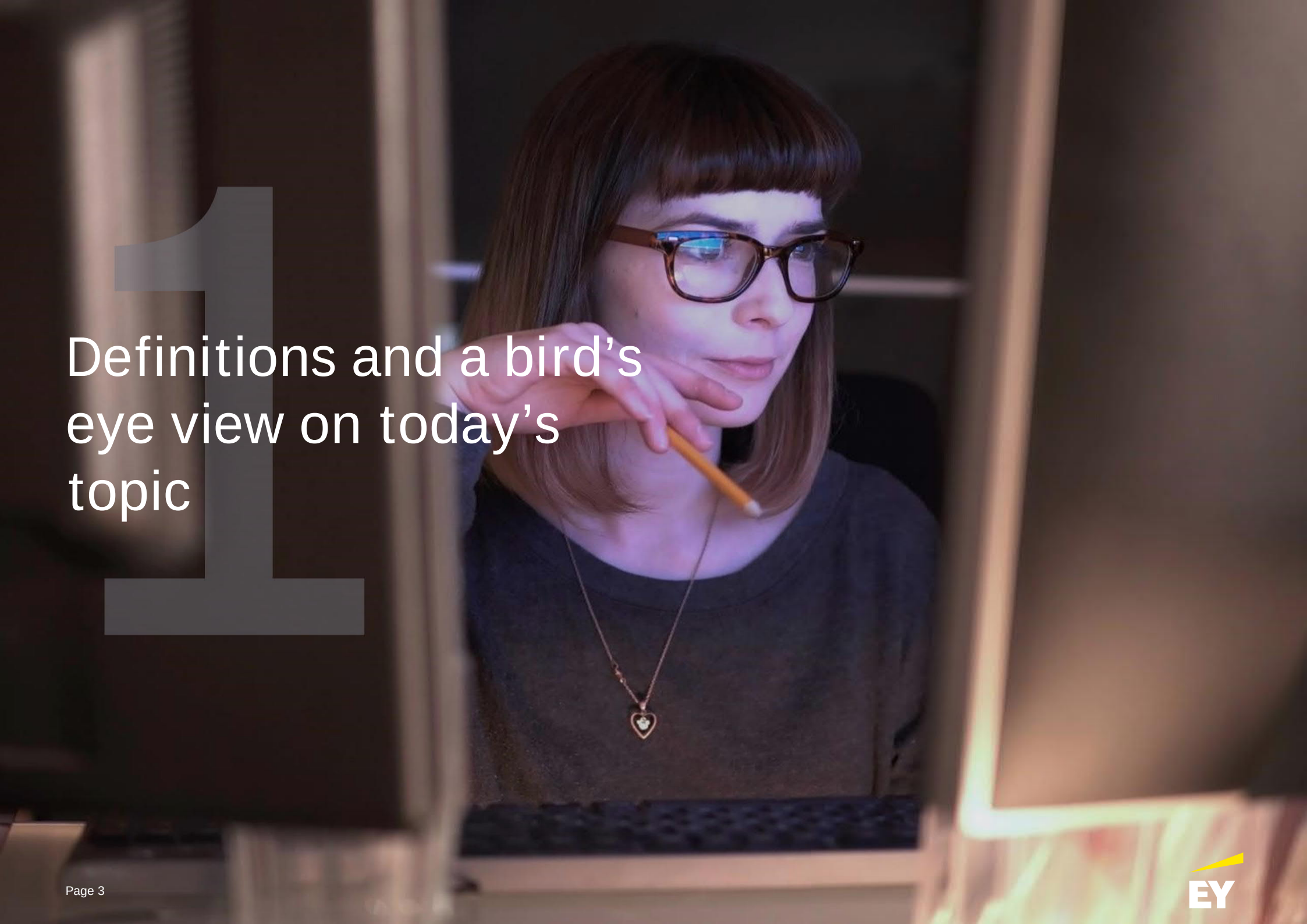
Here with you today:

Dr. Bernhard Hein

Partner EMEIA Financial Services - Advisory

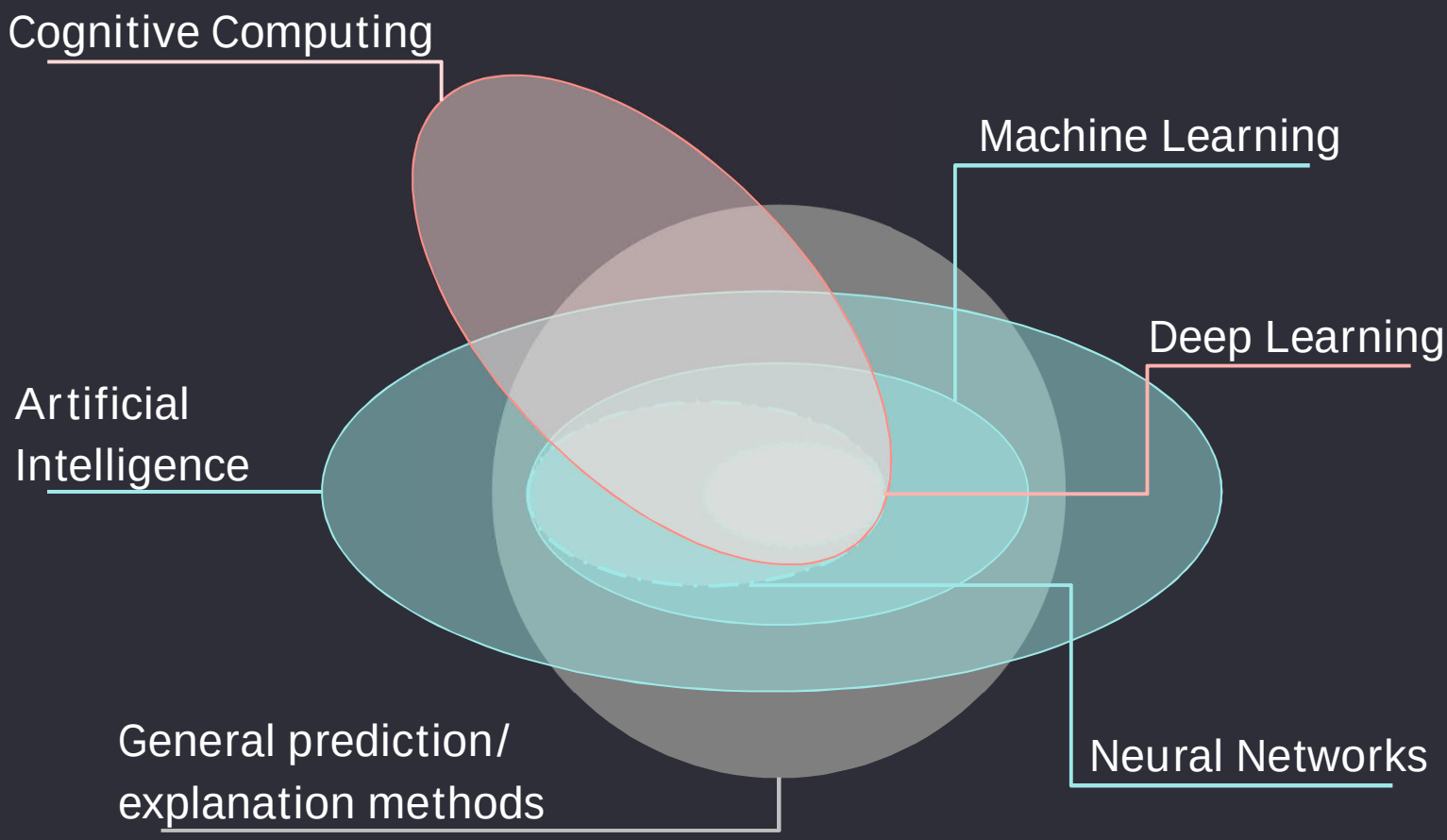
Dr. Ansgar Koene

Global AI Ethics and Regulatory Leader



1 Definitions and a bird’s eye view on today’s topic

A terminology convention for today (not overly precise)



Challenges we see

- Does an institution have a dedicated ML/AI definition, e.g. in their Model Risk Management Framework
- Is there clarity about the question where and to what degree additional MRM requirements are applicable for which models and their implementation

Within the realm of general prediction and explanation methods, we shall understand the notion of Artificial Intelligence to essentially contain all strictly algorithmic methods and exclude pure expert intuitive judgement / „single case opinions“.

Suggested convention, ML/AI starts where explainability becomes a challenge (regression is not ML)

Artificial intelligence (AI) is the theory and development of computer systems able to perform tasks that traditionally have required human intelligence[1]. It is broadly applied when a machine mimics cognitive functions that humans associate with other human minds, such as learning and problem solving.

Machine learning (ML) is a sub-category of artificial intelligence that gives computers the ability to learn without being explicitly programmed. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data with limited or no human intervention versus traditional algorithmic models used in financial services firms to identify casual relationships and patterns.

Supervised Learning

The algorithm is fed a set of 'training' data that contains labels on some portion of the observations, and it 'learns' a general rule of classification that it will use to predict the labels for the remaining observations in the data set

Unsupervised Learning

Data provided to the algorithm does not contain labels. The algorithm detects patterns in the data by identifying clusters of observations that depend on similar underlying characteristics

Reinforcement Learning

In between supervised and unsupervised learning. The algorithm is fed an unlabelled set of data, chooses an action for each data point, and receives feedback (perhaps from a human) that helps the algorithm learn

Range of machine learning techniques

Machine Learning Techniques	Data Requirements			Calibration Complexity			Intuitiveness			Model Performance		
	L	M	H	L	M	H	L	M	H	L	M	H
Linear Regression	●			●					●	●		
Logistic Regression		●			●				●	●		
GLM, GAM		●			●			●			●	
Decision Tree		●			●			●			●	
Naïve Bayes		●			●			●			●	
Bayesian network			●		●			●			●	
K-Nearest Neighbors		●		●				●			●	
Support Vector Machine		●			●			●				●
Ensemble Machine			●			●	●					●
Feed Forward Neural Networks			●			●	●					●
Recurrent Neural Network			●			●	●					●
Convolutional Neural Network			●			●	●					●

[1]Definitions from the FSB's report – 'Artificial intelligence and machine learning in financial services - Market developments and financial stability implications'

Complex ML techniques; higher priority for AI governance

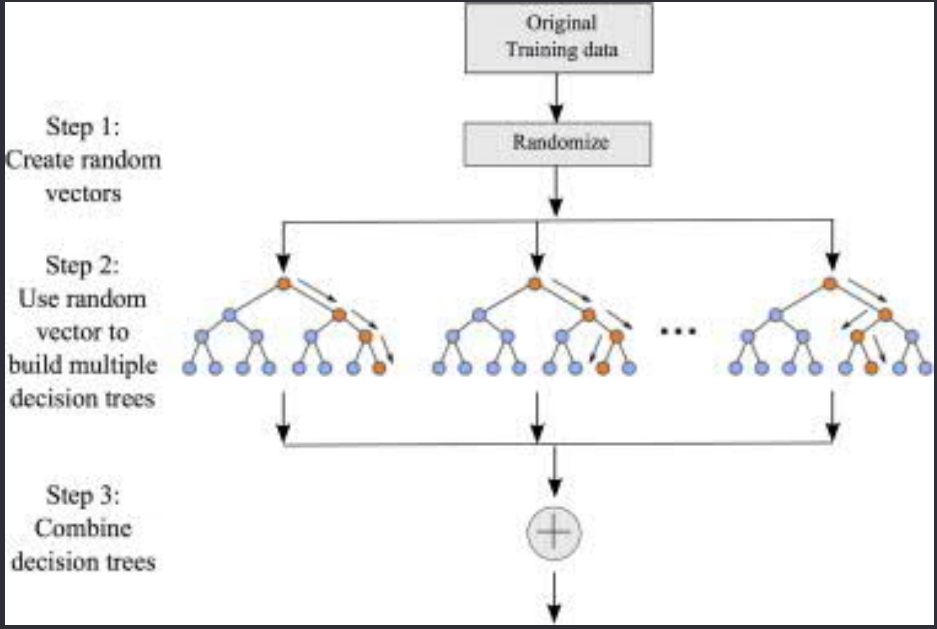
“Explainability becomes a challenge” means the generic model description does not tell you what the model actually does in a useful way

(part of a) Classical Regression model … and you can see what it does, and have some opinion on plausibility at once

Monthly financial freedom	
Category	Score
No information	0
Less than or equal to 0	22
Up to 500	-27
500 to 1000	14
1000 to 1500	38
1500 to 2000	48
More than 2000	48

Salary/total income	
Category	Score
No information	-1
Up to 87%	-31
87% to 94%	5
94% to 99%	26
Above 99%	23

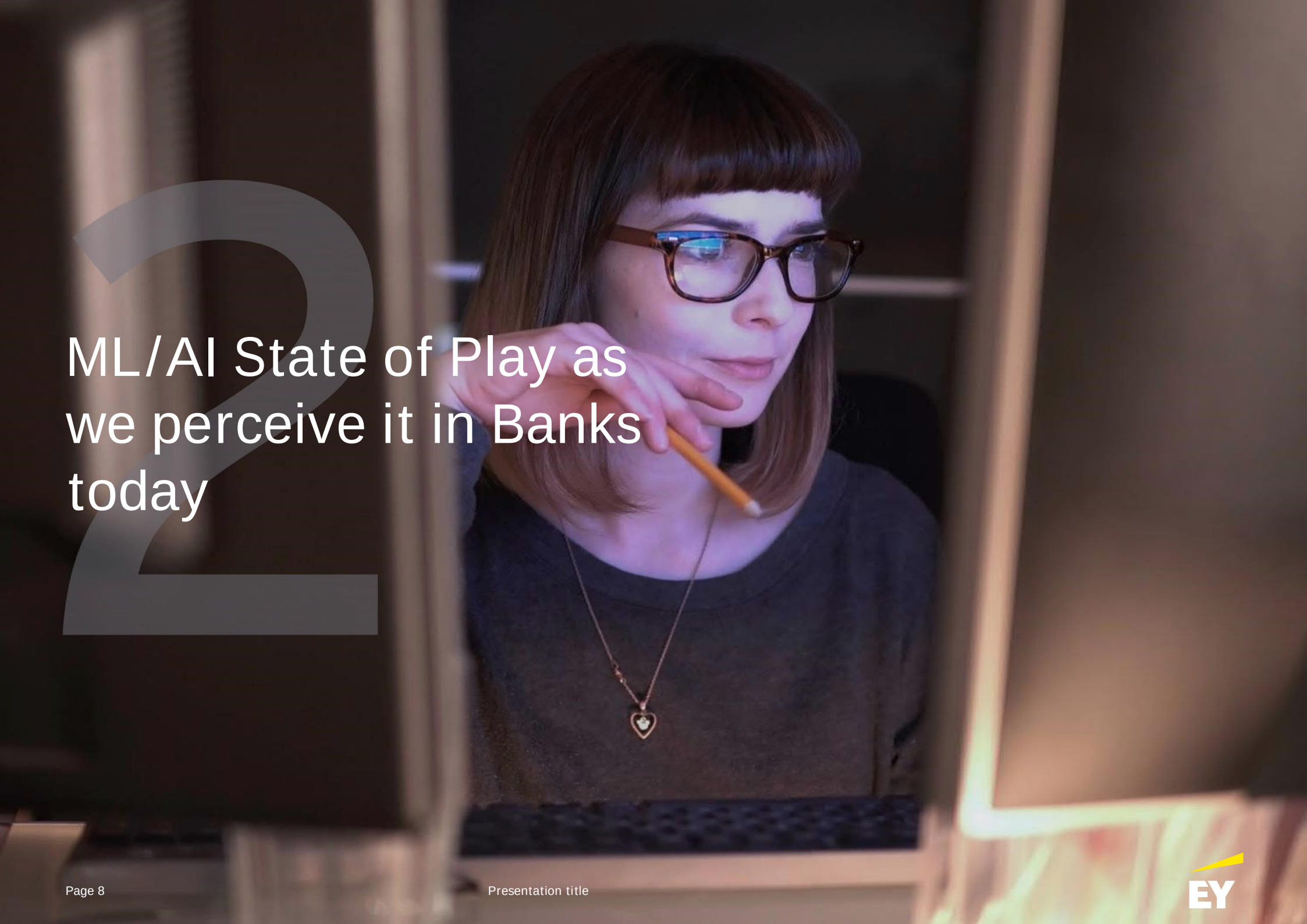
Random Forest … you will not be able to find such a description – a random forest prediction is a democratic vote of hundreds or thousands of decision trees, each of which has a relatively simple description … but to understand the forest, you need to understand every (.) tree. … can you?



Not everything changes with AI/ML instead of “classical” modelling

Remember, even if it’s ML/AI

- It’s still models
  - Need for clear governance, including definition, inventory, roles and accountabilities along the full model lifecycle – just as for any other model
- It’s still statistics
  - Data lineage, privacy, quality, sparsity, representativeness – not less of a challenge then in classical modelling
- It’s still a bank

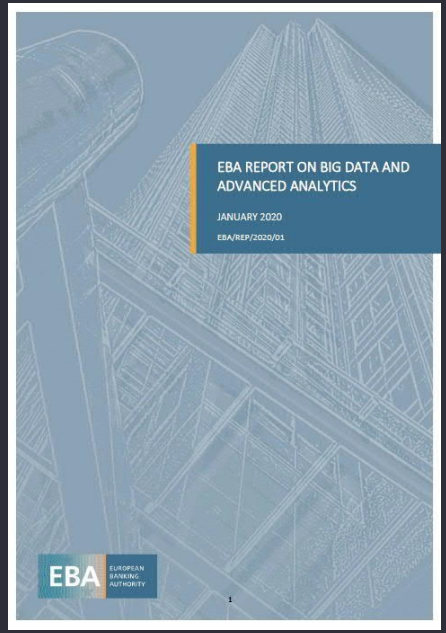


2

ML/AI State of Play as we perceive it in Banks today

EBA Report on Big Data and Advanced Analytics offers a very good overview that corresponds to our perception of the market

- 2 to 12 months required to move from early stage to production in an advanced analytics solution (EY: if IT is already capable)
- Institutions appear to focus on simple models for predictive analytics, prioritizing explainability and interpretability over accuracy and performance
- Legacy systems are an obstacle to adoption of AA, increasing reliance on cloud service providers to overcome this issue
- primarily internal data from core banking systems, rather than external data
- Preference for relatively simple algorithms (more explainable), to avoid black box issues (e.g. a preference for decision trees and random forests rather than deep learning techniques)
- some institutions share anonymized customer data with technology providers for model-training purposes. Some other institutions share their customer data with universities and public institutions only, and not actively with other commercial enterprises
- Some institutions address bias (sic!) at the model development stage by removing specific (sensitive) variables

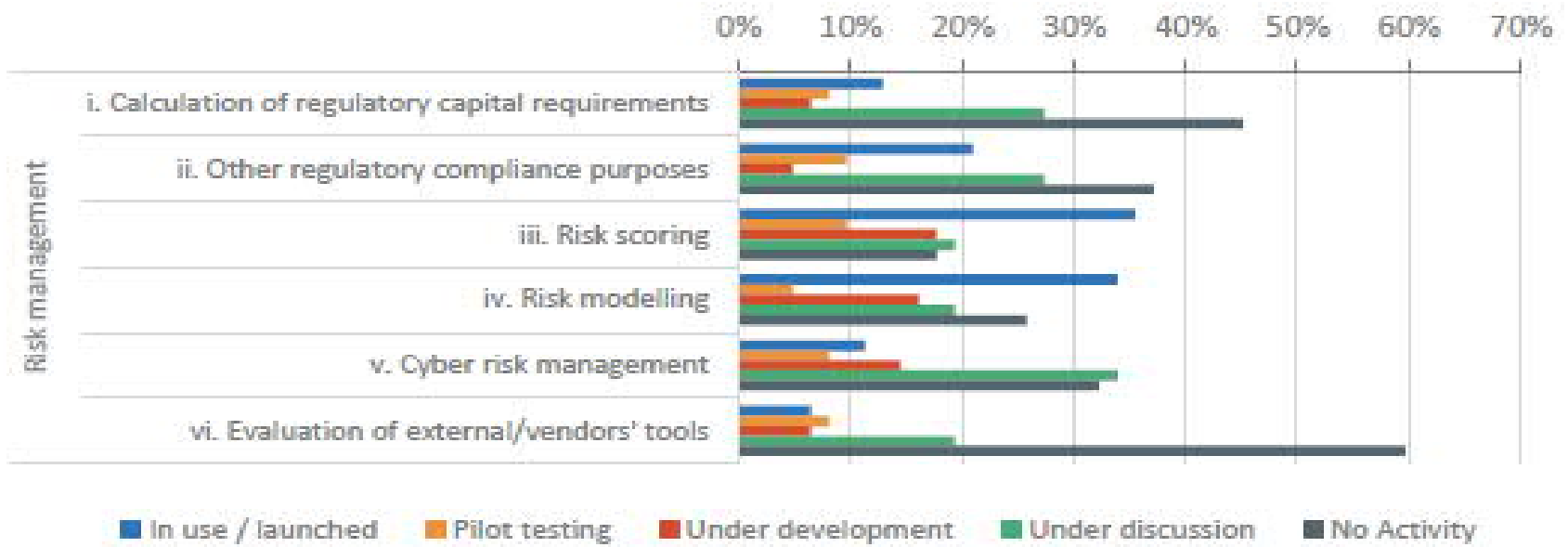


Reproducibility is sometimes challenging, but some risk-scoring and risk-modelling is observed

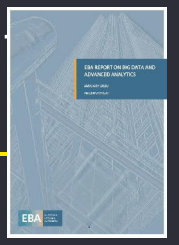
It appears that it is not always the case that the open source tools support the entire data science process that leads to a specific output in a reproducible way, as in some institutions only the source code is recoverable while in other institutions all relevant events are reproducible.

Growing use was observed for risk-scoring and risk-modelling purposes:

Figure 2.1: Current use of Big Data Analytics for risk management purposes

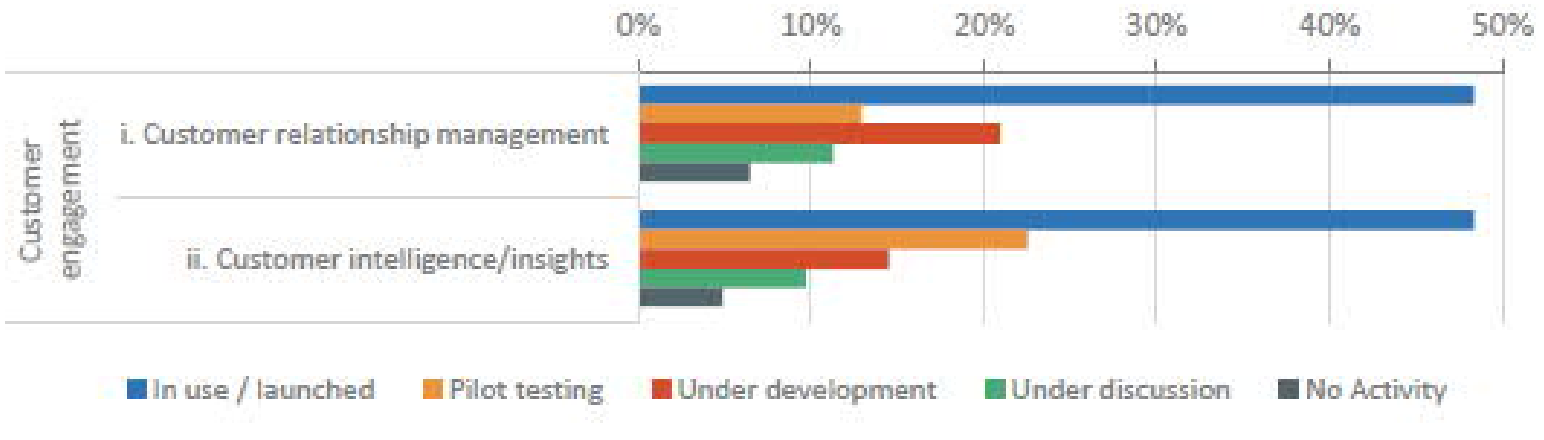


The current landscape still largely avoids the prudential domain, and that's for good reason, says EBA report



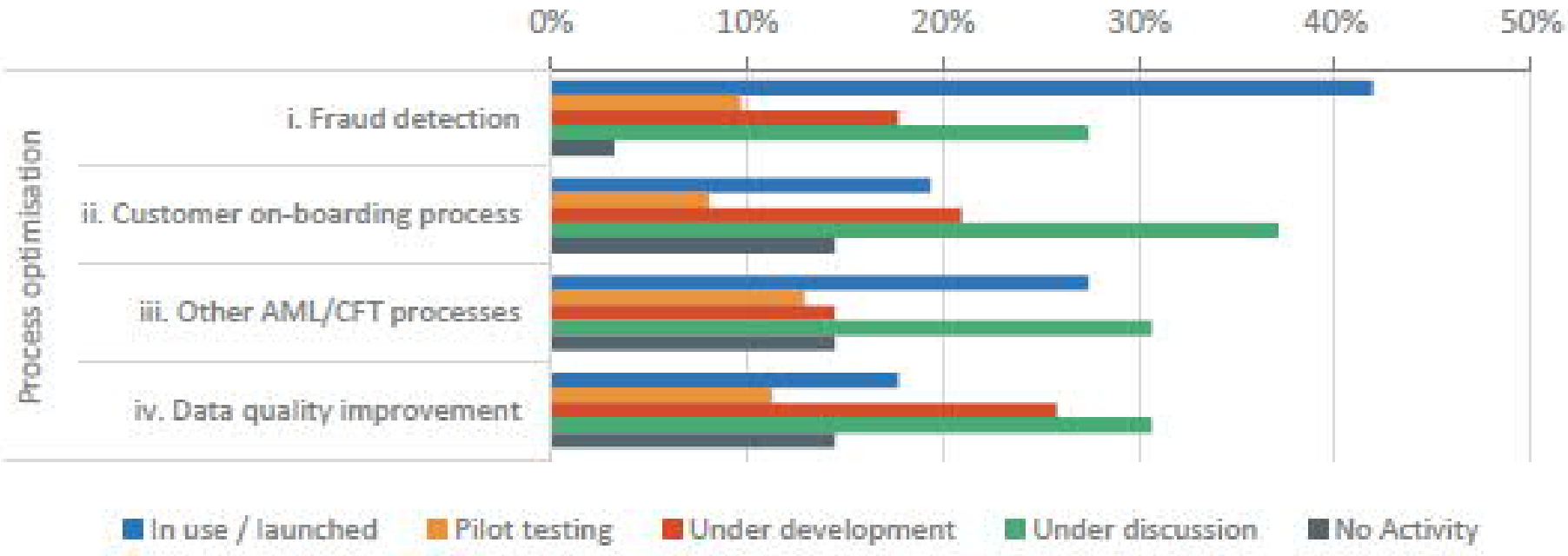
- Fraud detection (real time) is a major focus
- Institutions are increasingly turning to additional data sources, unstructured and semi-structured, including on social media activity, mobile phone use and text message activity, to capture a more accurate view of creditworthiness. E.g. timely payment of utility bills for individuals without credit history
- From a prudential framework perspective, it is premature to consider ML an appropriate tool for determining capital requirements, taking into account the current limitations
- Significant use of (e.g. NLP enabled) AI for customer interaction, less on market analysis (e.g. product pricing):

Figure 2.2: Current use of Big Data Analytics for customer engagement purposes



AI/ML are already a lot stronger in process optimization than in other areas

Figure 2.4: Current use of Big Data Analytics for process optimisation purposes



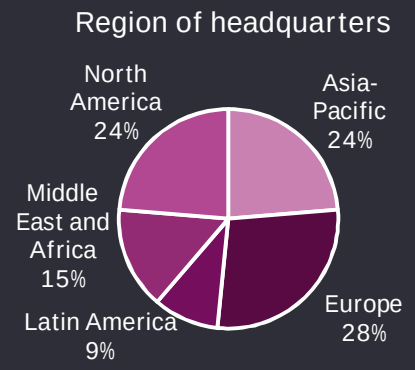
We see very similar overall results in our tenth annual EY/IIF global bank risk management survey – see following slides

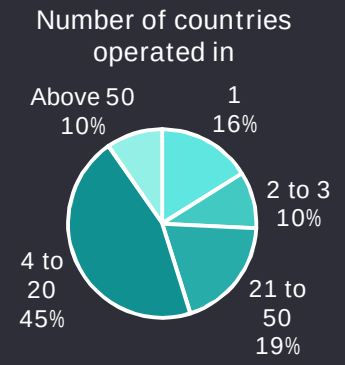
Research methodology and participant demographics

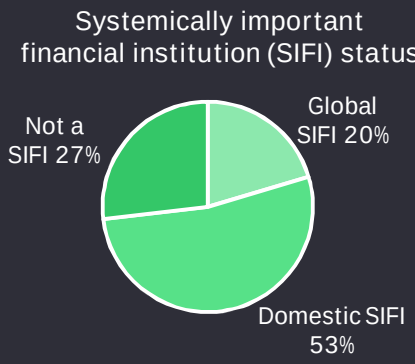
EY, in conjunction with the IIF, surveyed IIF member firms and other top banks in each region globally (including a small number of material subsidiaries that are top-five banks in their home countries) from June 2019 through Sept 2019. Participating banks' CROs or other senior risk executives were interviewed, completed a survey, or both.

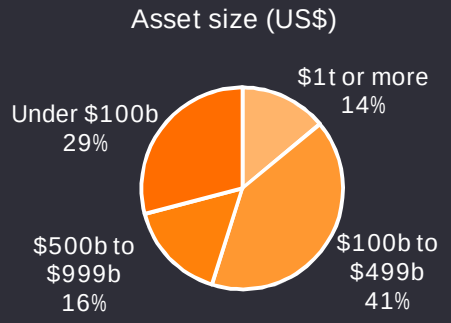
In total, 94 firms across 43 countries participated (up from 74 banks in 2018). Regionally, those banks were headquartered in Asia-Pacific (21), Europe (26), Middle East and Africa (14), Latin America (10) and North America (23). Of those, 19 are globally systemically important banks and 49 have been designated as systemically important domestically. Data in this report relate to the 92 banks that completed the quantitative survey, and the narrative includes insights gleaned from qualitative interviews with some of those and other banks. Participating banks were fairly diverse in terms of asset size, geographic reach and type of bank.

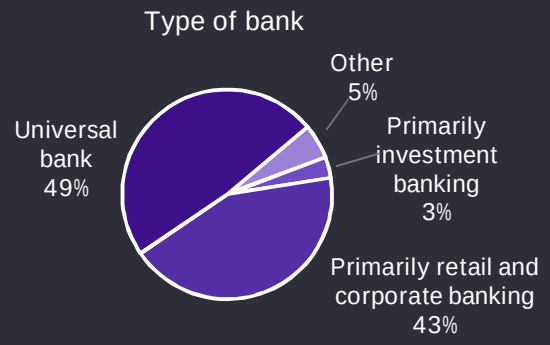
Note that 21 additional financial institutions participated informally by responding to the survey. Their data are not included in this survey report, but directionally they did inform the narrative.











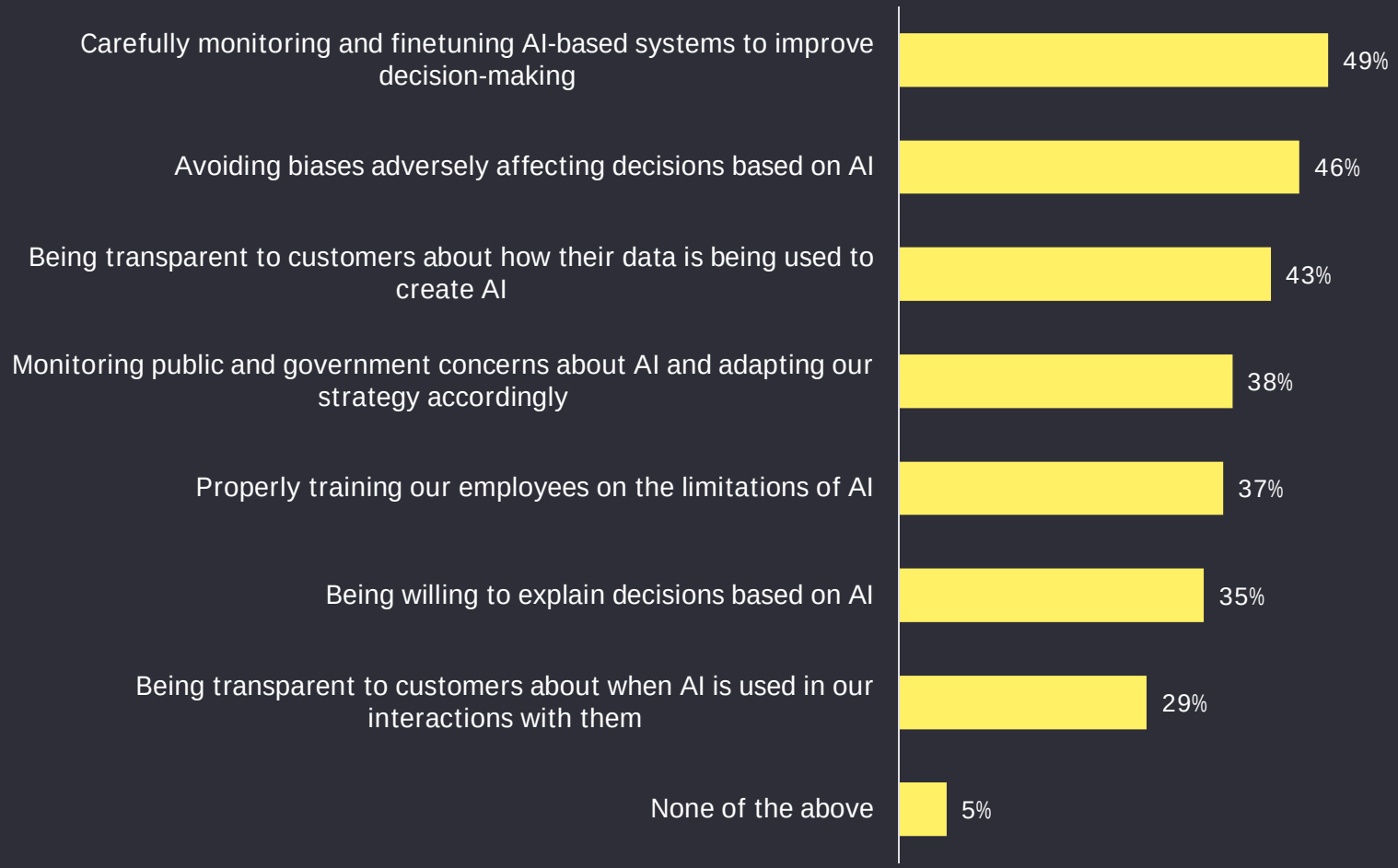
CROs expect key areas for AI/ML Usage in banks to saturate at around 70% adoption rate across use cases

areas will change over the next five years as ML/AI gets used across all 3 LoD, but not for everything and not by everyone.

Multiple responses allowed

Making better decisions, being fair and transparent are most important for banks to solve ethical challenges around ML/AI use

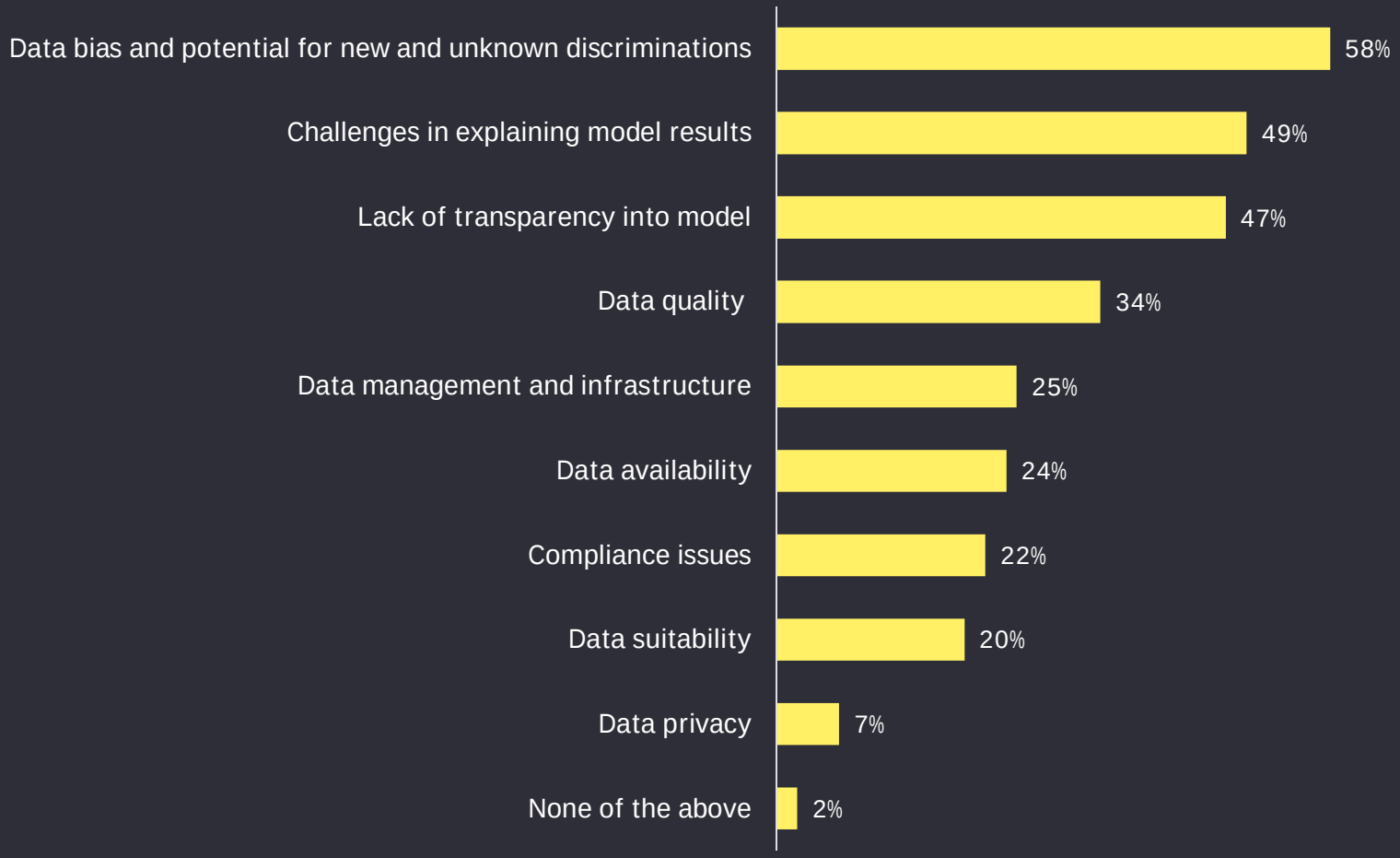
What are the top three ways of addressing public or customer concerns about potential ethical issues related to the use of artificial intelligence?



Multiple responses allowed

Correspondingly, avoiding bias and coming to terms with model explainability and transparency figure highest as MRM concerns

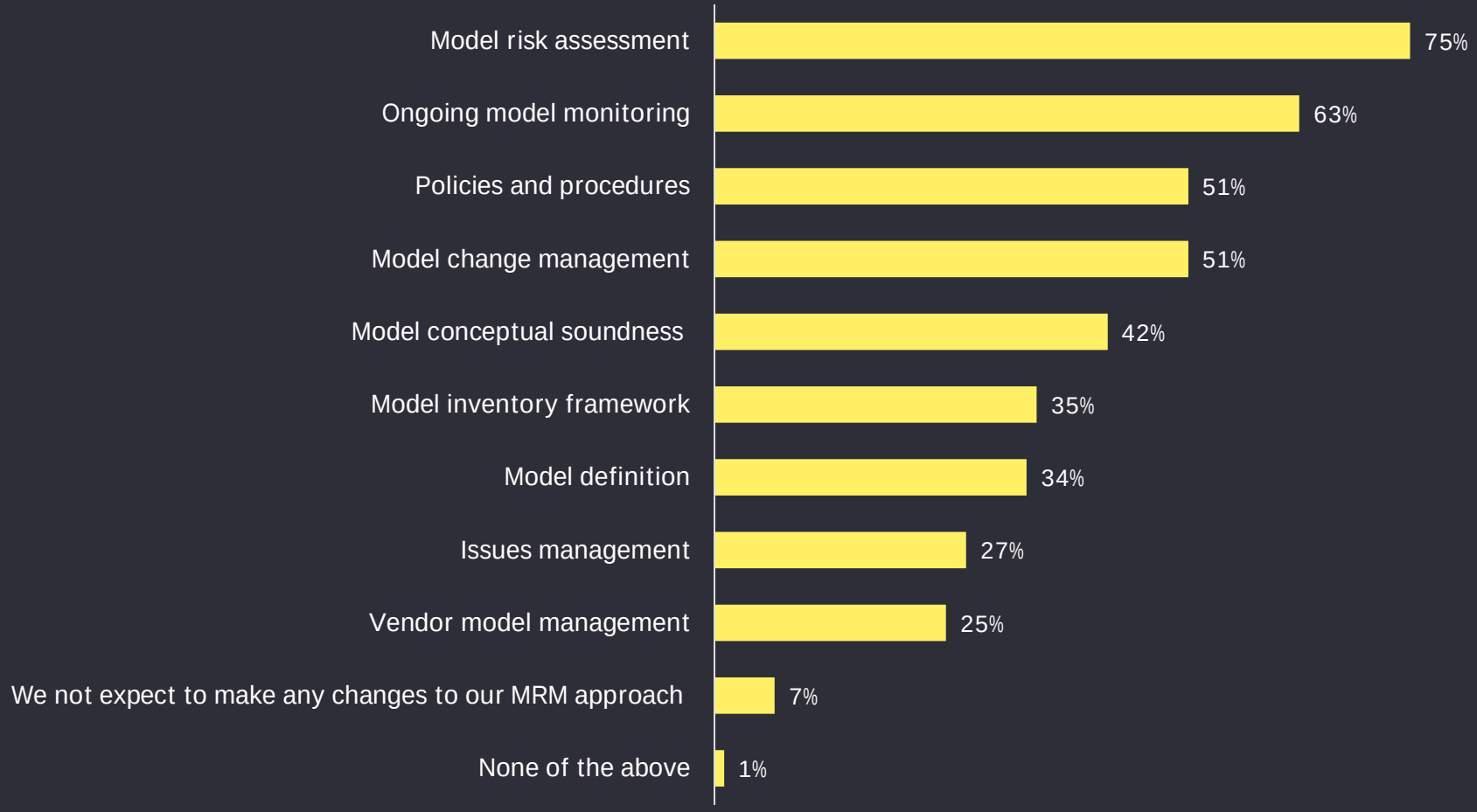
From a model risk management (MRM) perspective, what are your top three concerns about using artificial intelligence?



Multiple responses allowed

Expected MRM enhancements center around better assessment of model risk, model monitoring and governance topics

Over the next three years, in which of the following areas do you intend to enhance your model risk management (MRM) process to adapt to governance requirements for AI/ML?

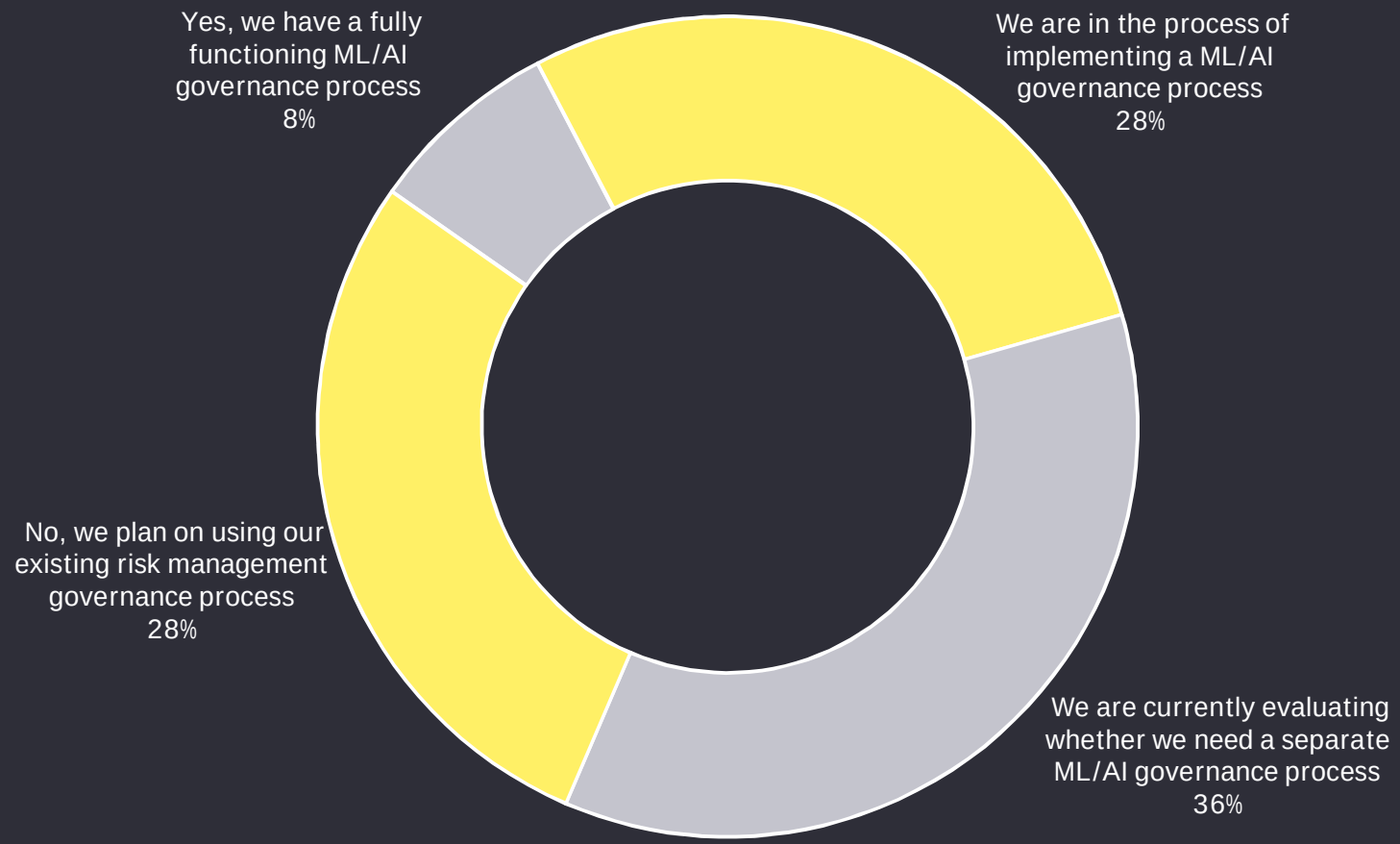


Multiple responses allowed

Distinct governance structures for ML/AI seems lagging behind to some extent

- Few banks have yet built a distinct governance approach to ML and AI
- Many others will follow suit in the next few years

Does your organization have a distinct governance structure for the use of machine learning (ML) and/or artificial intelligence (AI)?

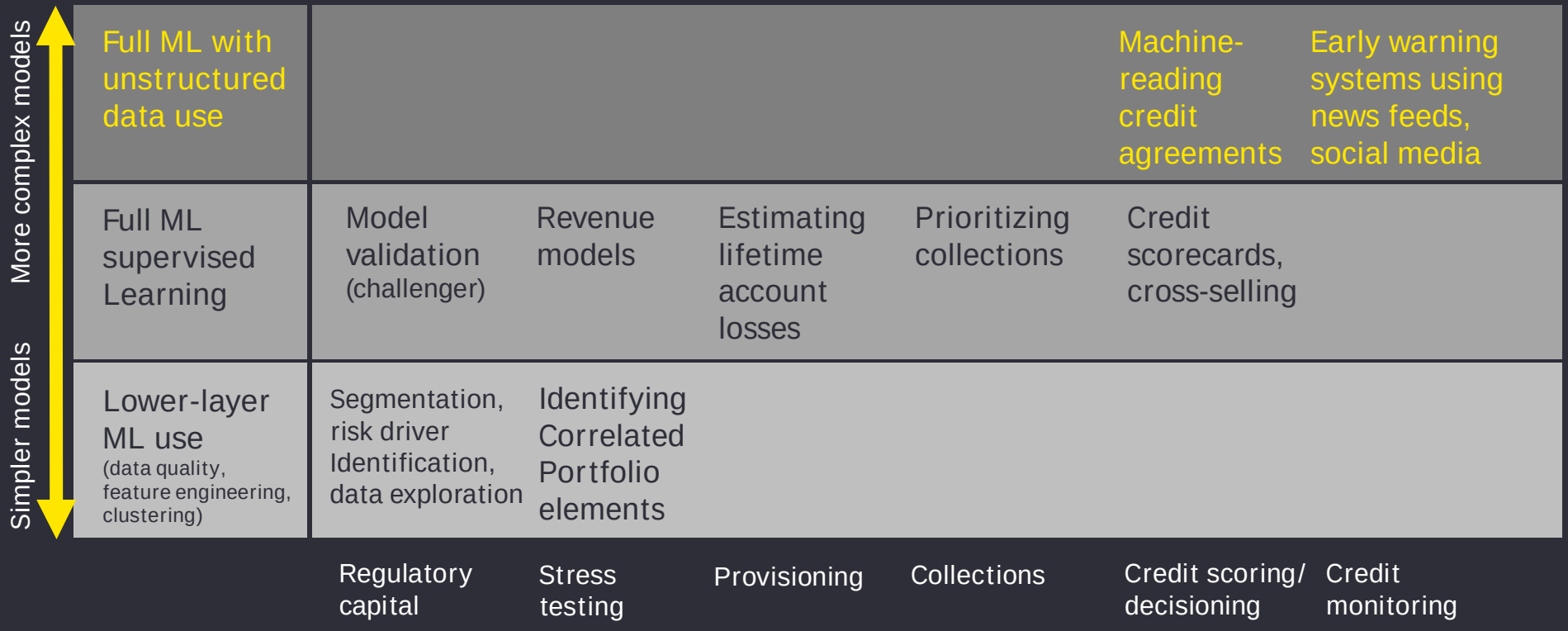


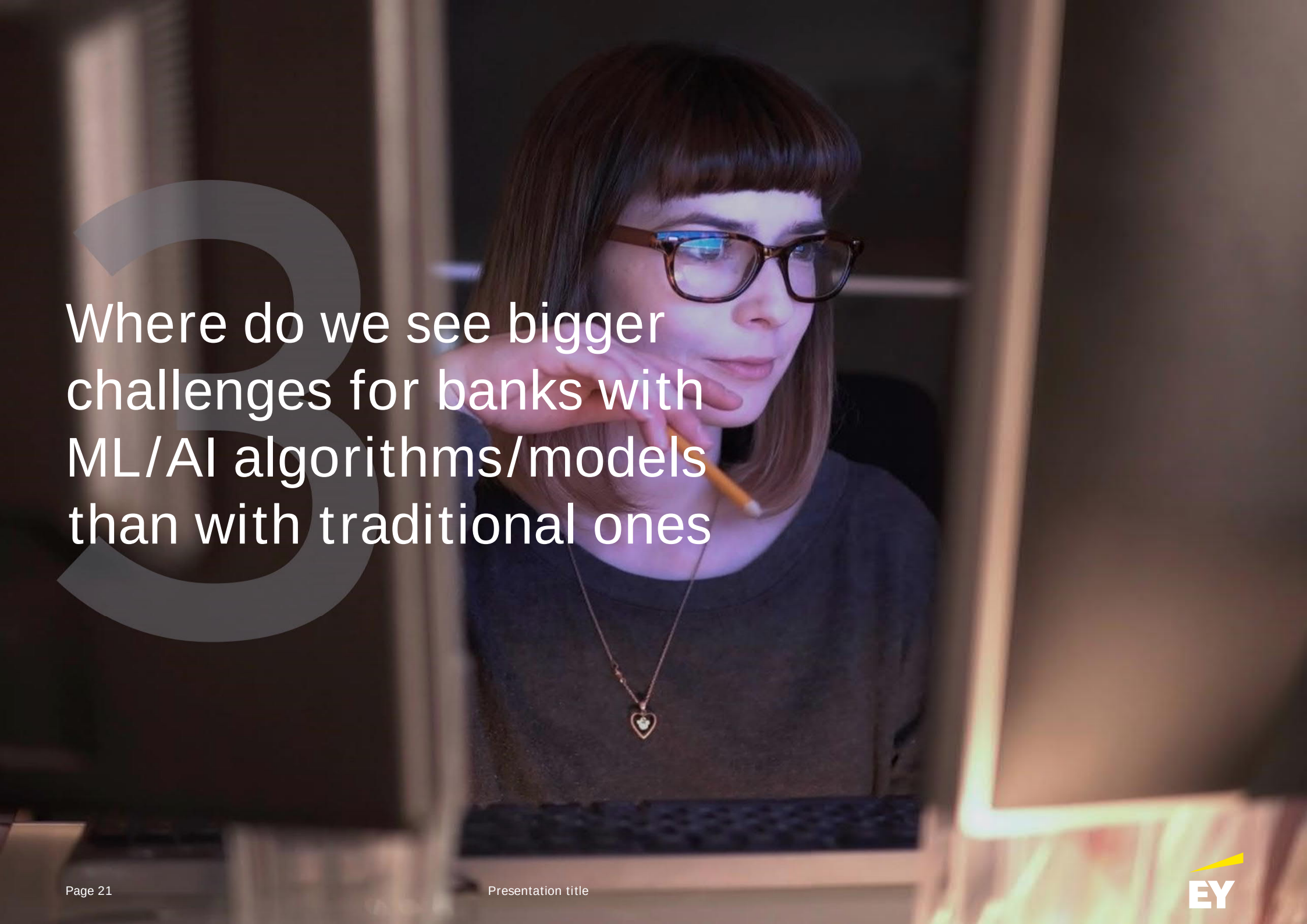
Second-line risk role has some development potential if 2nd line wants to be “on top” of ML/AI as much as for models from the prudential space

- Yet: limited role in managing risks linked to the use of ML/AI
- Will likely change over the next few years

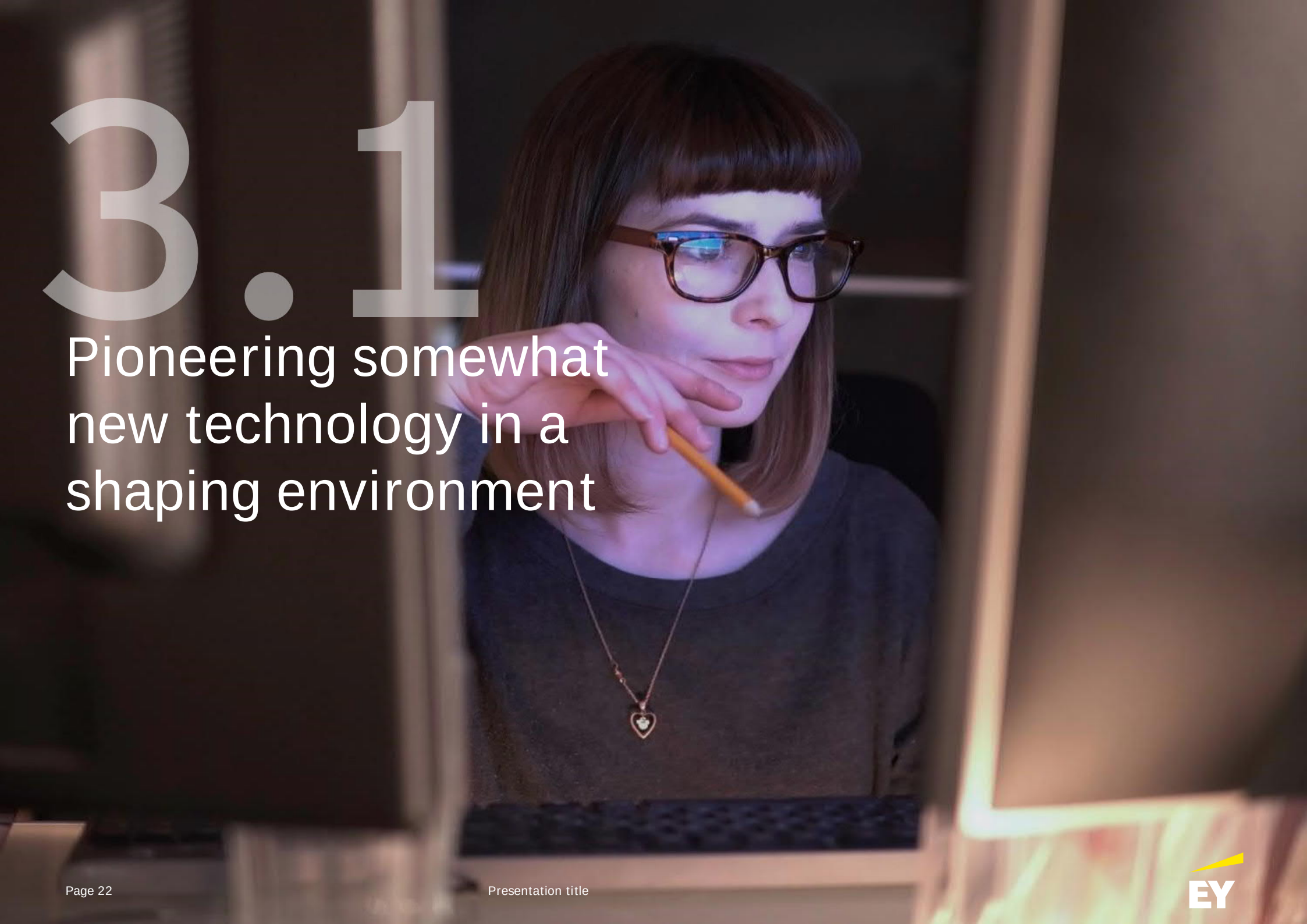
Departing from the IIF/EY study, we generally see a (small) bit more reluctance in the European Market than globally

Our somewhat more local impressions from SSM markets (and some others) see more reluctance around AU/ML use in the core supervised domains. Banks are gaining experience in the periphery, and where automated decisions have less visibility to the public, and hence cause less reputational risk.





Where do we see bigger challenges for banks with ML/AI algorithms/models than with traditional ones



3.1 Pioneering somewhat new technology in a shaping environment

Rising global push for guidelines on trustworthy and ethical use of AI and data

AI Ethical Principles – high levels of consensus

- EC High Level Expert Group for AI Ethics Guidelines for Trustworthy AI
- OECD Recommendations on Artificial Intelligence (May 2019)
- G20 Principles for responsible stewardship of trustworthy AI (endorsed by leaders 06/2019)
  1. Inclusive growth, sustainable development and well-being
  2. Human-centred values and fairness
  3. Transparency and explainability
  4. Robustness, security and safety
  5. Accountability
- OECD AI policy observatory
  (launch 27 February 2020)

Cautious approach to translating AI Principles into Legislation

EU

- Identify gaps where regulation of AI is not already covered by existing general or sectoral legislation, e.g. Banking sector Model Risk Management regulations, GDPR.
- Risk and Impact assessment dependent AI specific regulatory requirements to be considered through public consultation.

US

White House instruction to government agencies to:

- pursue a risks based approach and avoid a precautionary principle based approach to AI regulation;
- avoid taking regulatory action unless there is clear evidence that the harms a particular AI outweigh the benefits.

Encouraging the development of Standards for AI systems

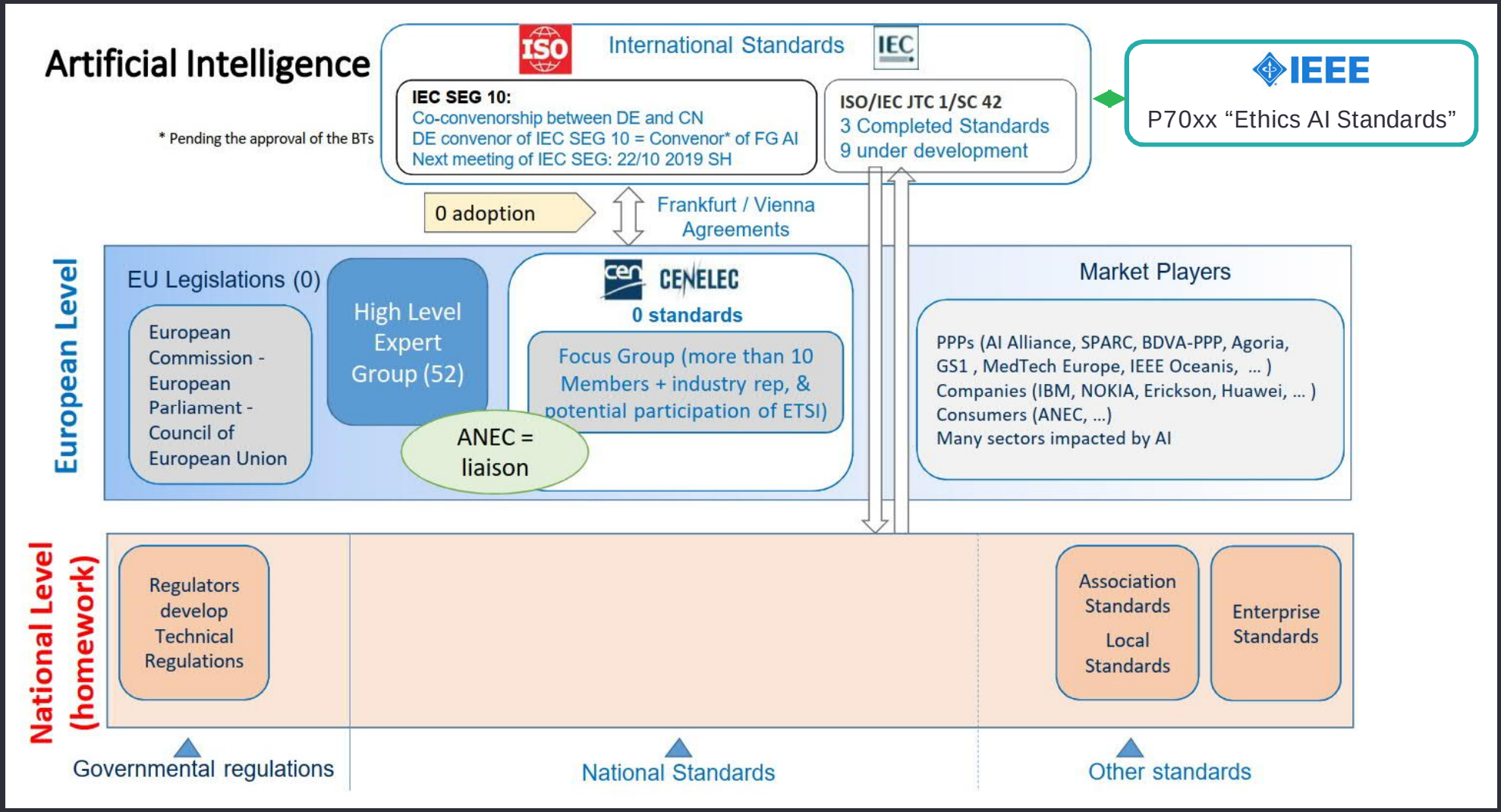
EU (starting mid 2019)

Coordination between the EC (DG CNECT, DG Just, DG Grow) and the European Standards development bodies CEN-Cenelec and ETSI to survey and identify specific standardisation needs for the EU.

US (February 2019)

Executive Order 13859 on Maintaining American Leadership in Aritficial Intelligence instructed NIST to develop a “plan for federal engagement in developing technical standards and related tools”.

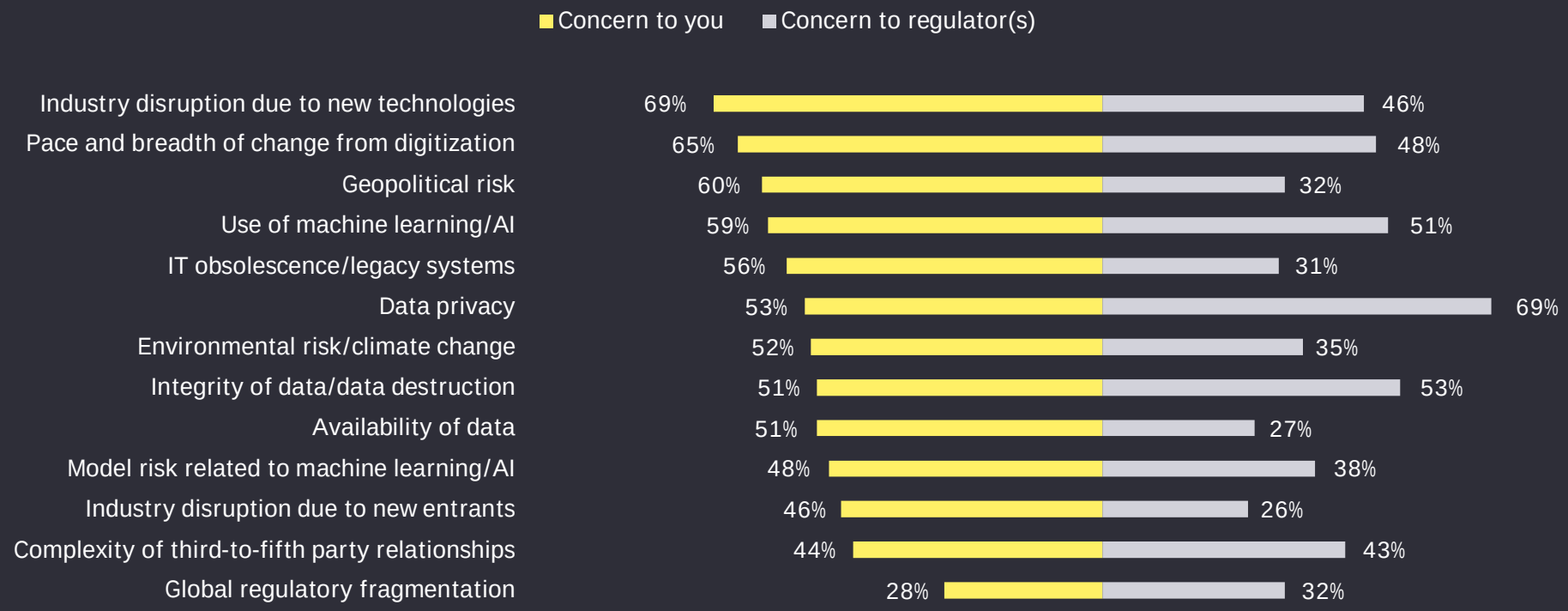
We believe that some conceptual work needs to be done by Standard setters, upon which FS regulation could build



Who wants to be the first to explain all this ML/AI stuff to … (you? Internal audit? The statutory auditor?)

Technology change – including AI and legacy systems – is a major emerging risk, and acknowledged as major regulatory concern – likely with pioneering effects on top of regulatory workload on model development and model validation units.

What emerging risks do you believe will be most important for your organization and your regulators* over the next five years?



Multiple responses allowed

*Represents banks' views on regulators' priorities; not regulators' views

The specialist skillsets required in risk around ML/AI are high in demand also in other industries, hence scarce

Banks highlight a set of mainly nonfinancial risk domains where they believe they still need to add specialists. For the desired profiles, the labour markets are small and there's a high demand, and banks compete strongly with other industries.

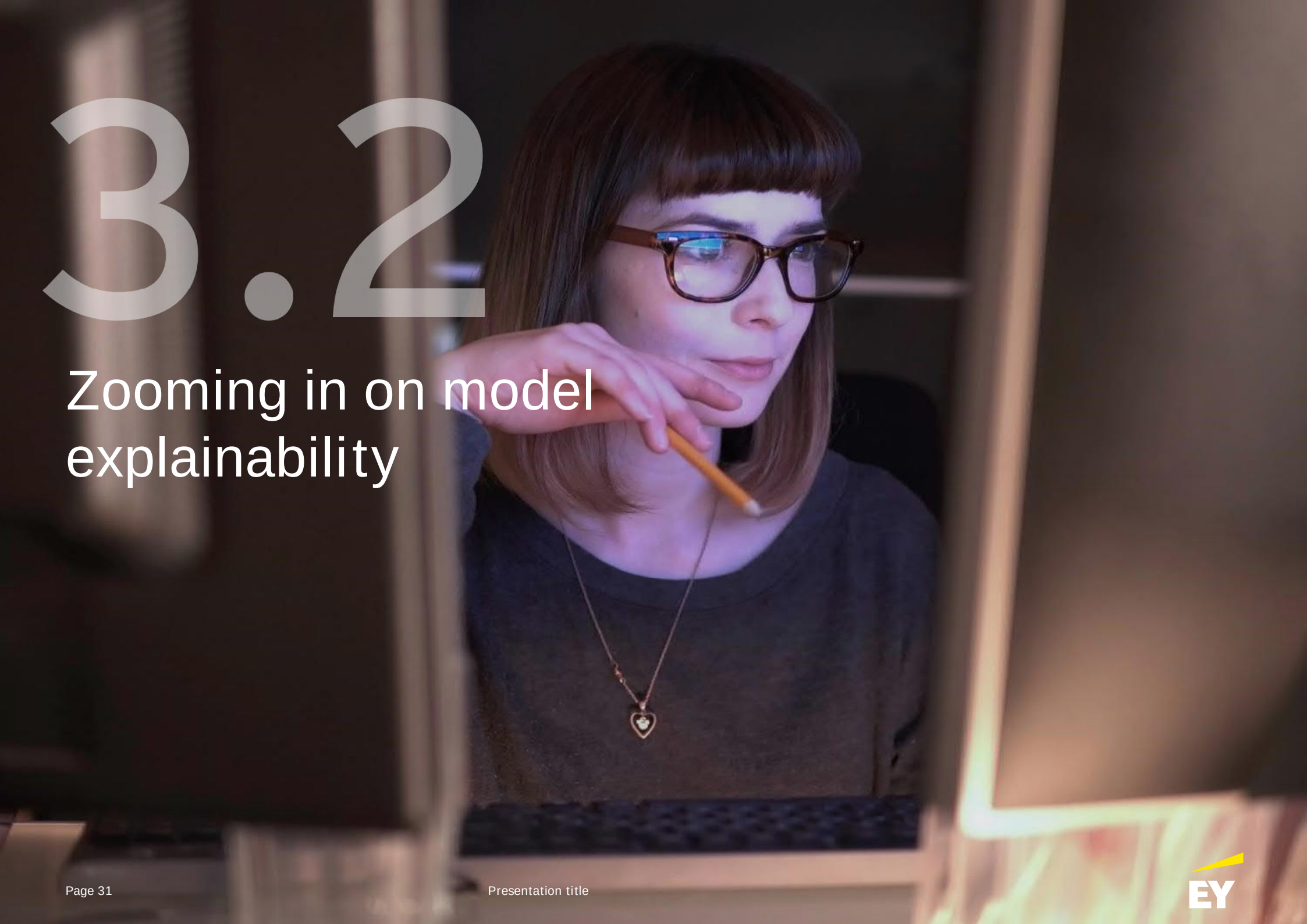
In which of the following risk areas are specialized talent or skill sets currently needed to better manage risks?

Multiple responses allowed

And then you have the model but your IT cannot easily implement …

While we have seen some model development units build prototypes, quite a few of these prototypes have not made it into production due to relatively tough IT hurdles:

- The description of a score card (logistic regression, typical scoring model) looks like just another simple Excel table and can be implemented at arms length
- The description of a random forest or a neural network is more unwieldy and it is an advantage to have production implementation in a very different way from development/sand-box, using different programming languages or code packages
- However, some open source packages used for ML model development are not easily suitable for a production quality implementation, which creates a tech hurdle
- This new technology requirement poses a hurdle that sometimes makes to more attractive for banks to do calculations in pre-equipped cloud environments

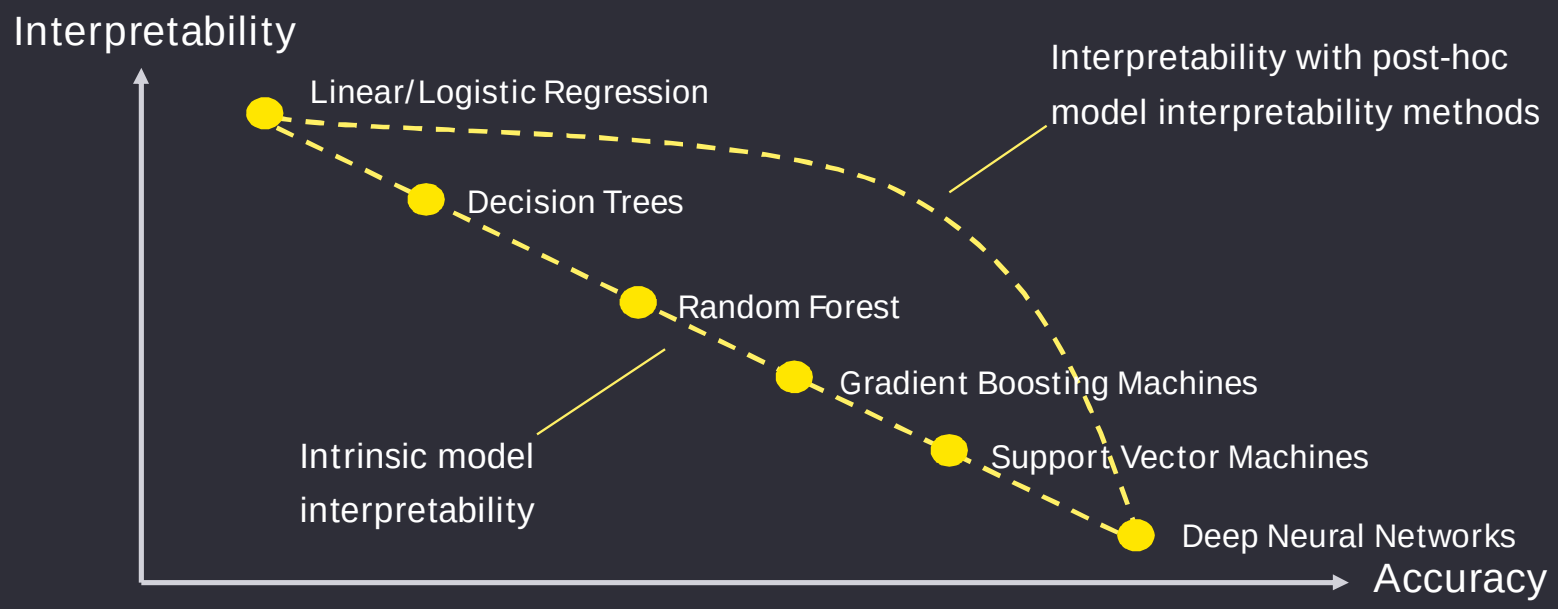


3.2

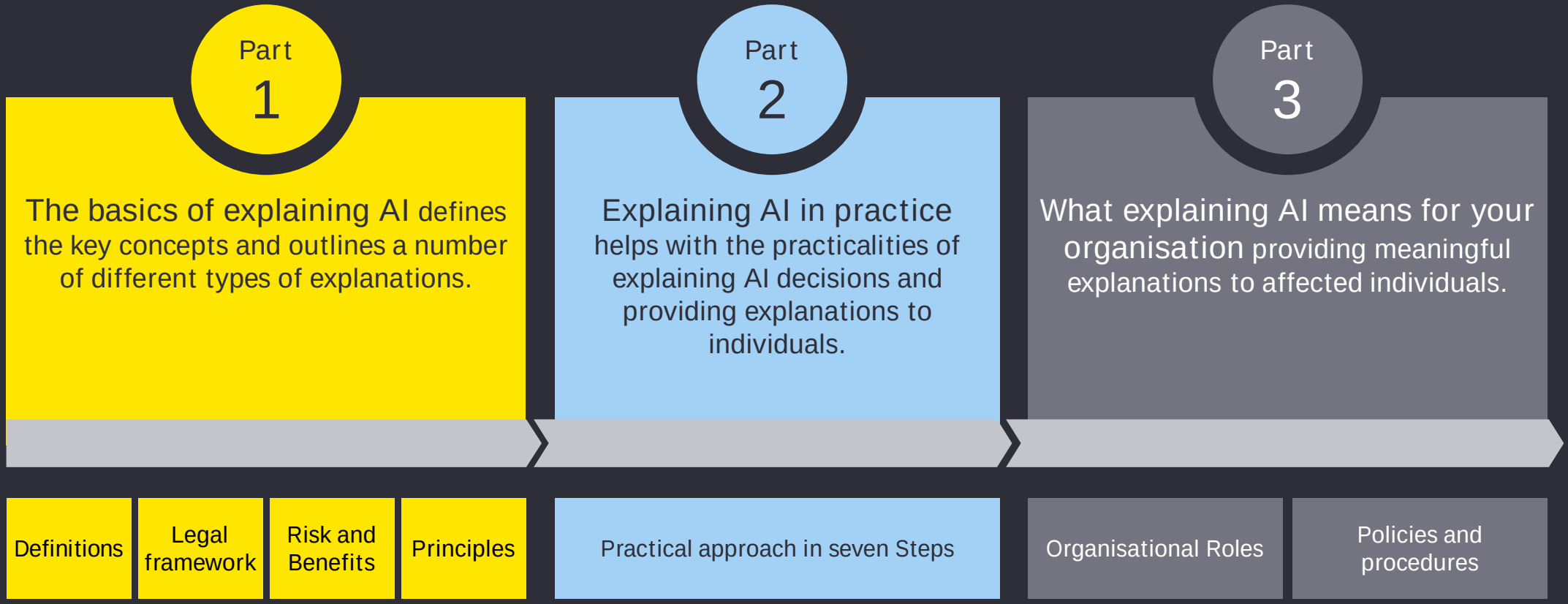
Zooming in on model explainability

With the currently available landscape of typical model designs, model accuracy and model interpretability are a trade-off

- ML increasingly used in settings where decisions strongly impact human life and business performance. Hence, critical to understand why these models make certain decisions.
- For many ML models, the relation between model input and output is difficult for humans to understand. Therefore, the models are said to have low interpretability (black boxes)
- With larger datasets and increasing importance of unstructured data, however, the best accuracy is often achieved by using complex models that are difficult to interpret.
- Consequently, increasing tension between model accuracy and model interpretability.



UK Data Protection Authority (ICO) draft guidance on ”Explaining decisions made with AI” explains different audiences have different explainability interest



Rational explanation: reason which led to the decision (non-technical way)

Responsible explanation: who is involved in the development, management and implementation and who to contact for human review

Data explanation: what data has been used and how; what data has been trained and tested

Fairness explanation: steps across the design and implementation of an AI system to ensure its decisions are unbiased and fair

Safety and performance explanation: design and implementation steps to maximise accuracy, reliability, security and robustness

Impact explanation: impact that the use of an AI system and its decisions has or may have as on individual

But then, the explainabtiliy challenge needs to be solved for a wealth of stakeholders, not only the public, depending on outside model visibility

BANK OF ENGLAND

Staff Working Paper No. 816

Machine learning explainability in finance: an application to default risk analysis

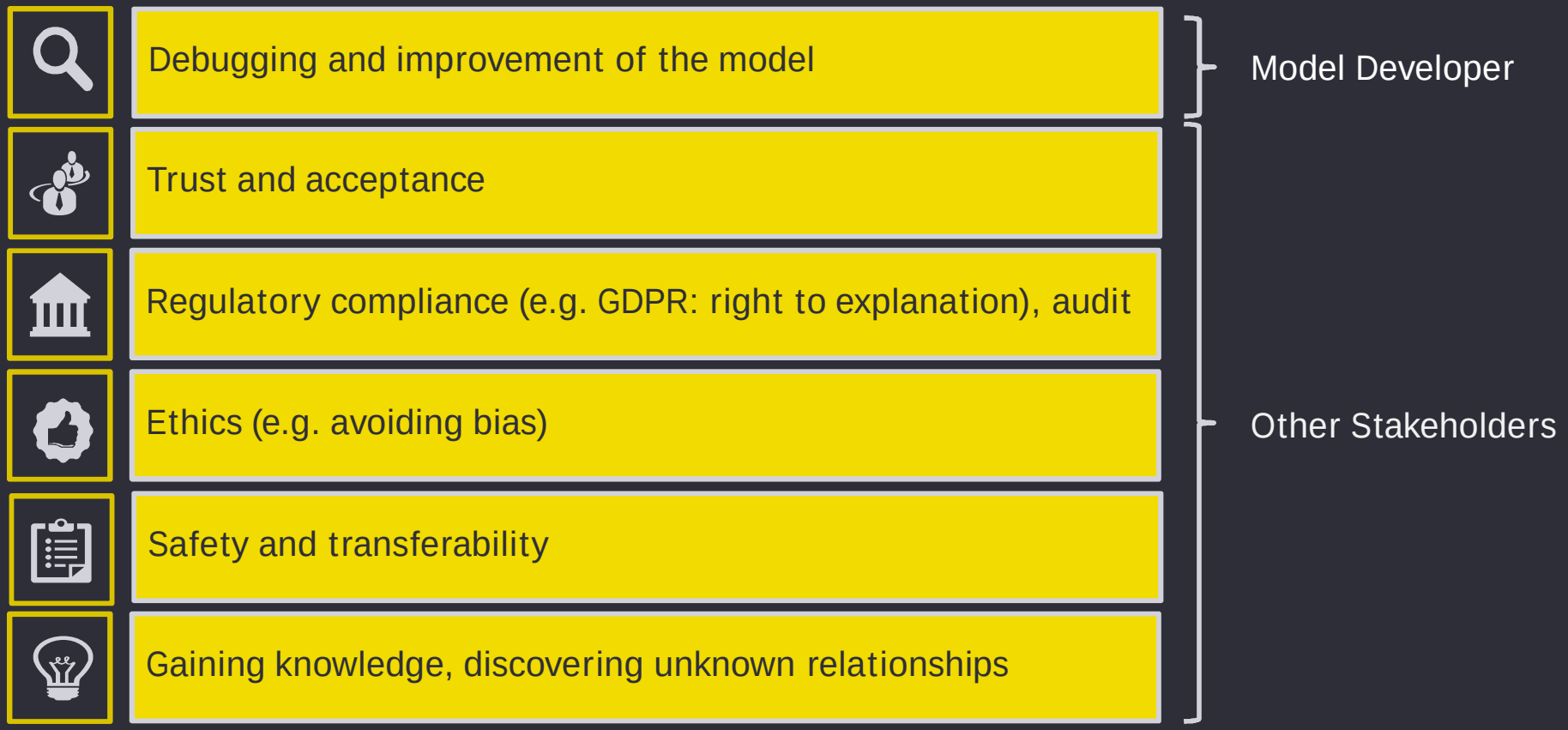
Philippe Bracke, Anupam Datta, Carsten Jung and Shayak Sen

	Stakeholder interest					
	Developer	1st line model checking	Manage-ment	2nd line model checking	Conduct regulator	Prudential regulator
1) Which features mattered in individual predictions?	X				X	
2) What drove the actual predictions more generally?	X	X	X		X	
3) What are the differences between the ML model and a linear one?	X	X				
4) How does the ML model work?	X	X	X	X	X	X
5) How will the model perform under new states of the world? (that aren't captured in the training data)	X	X	X	X	X	X

On the other hand, there are a lot more objectives that interpretability serves than just reaching a good model performance

During development, achieving best performance on test set is often the only objective

production requirements much more diverse and frequently include model interpretability



Talking about interpretability of (ML) models

Interpretability is split into:

1) model development process, 2) Global and 3) Local interpretability

1

Algorithm

How does the algorithm create the model?

Model

How does the model make predictions?

- Global model interpretability

Prediction

Why does the model make a specific prediction for one instance?

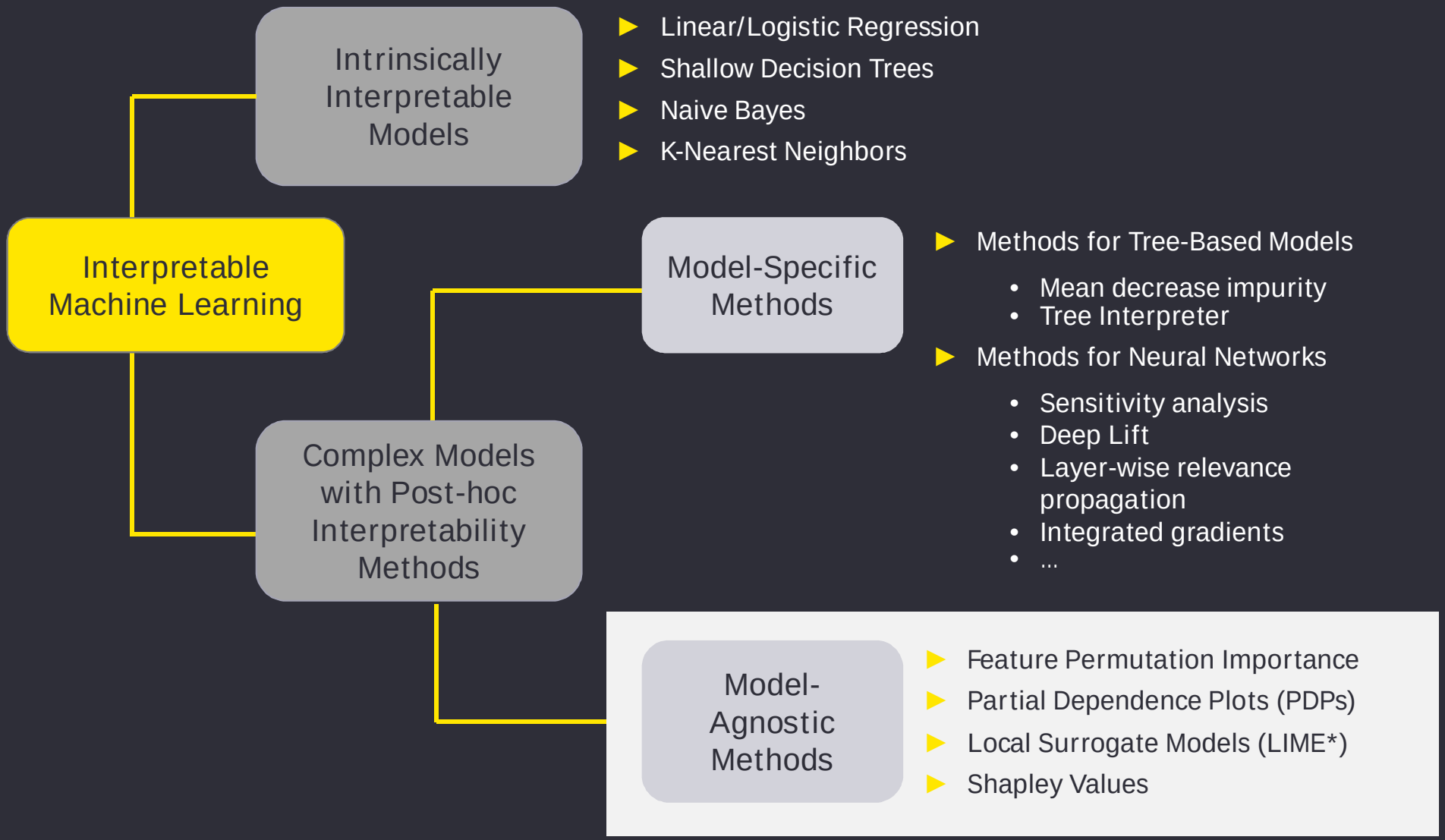
- Local model interpretability

2

Machine Learning model interpretability

- Intrinsic (ante-hoc) model interpretability
- Model interpretability with post-hoc interpretability methods (using interpretability methods that are applied after the model has been trained or after a prediction has been made)
  - Model-specific methods
  - Model-agnostic methods

Machine Learning interpretability can be tackled with a set of different methods depending on the choice of the algorithm (a focus on day 2 of this workshop)



A combination of several methods is regarded as giving the most intuitive answers to answer the 5 questions raised by Bank of England WP 816

A three step methodology is used in WP 816 – do not rely on a single recipe:

Quantitative Input Influence (QII)

Measures feature influences by intervening inputs and estimating their Shapley values

Clustering

Group of explanations for different areas of input space

Simulations

Conduct simulations on data that the model has not been trained or tested on

Further methods exist that can be used in combination to increase explainability:

Surrogate model

Construct surrogate model that approximates the workings of the complex ML model for all input data

Feature importance

Estimation of how much the model prediction variance changes due to exclusion of individual features. It does not capture feature interactions.

LIME method

Example based explanations which try to explain aspects of a model based on how it classified selected input regions

Instance based explanations

Explanations on a prediction-by-prediction basis.

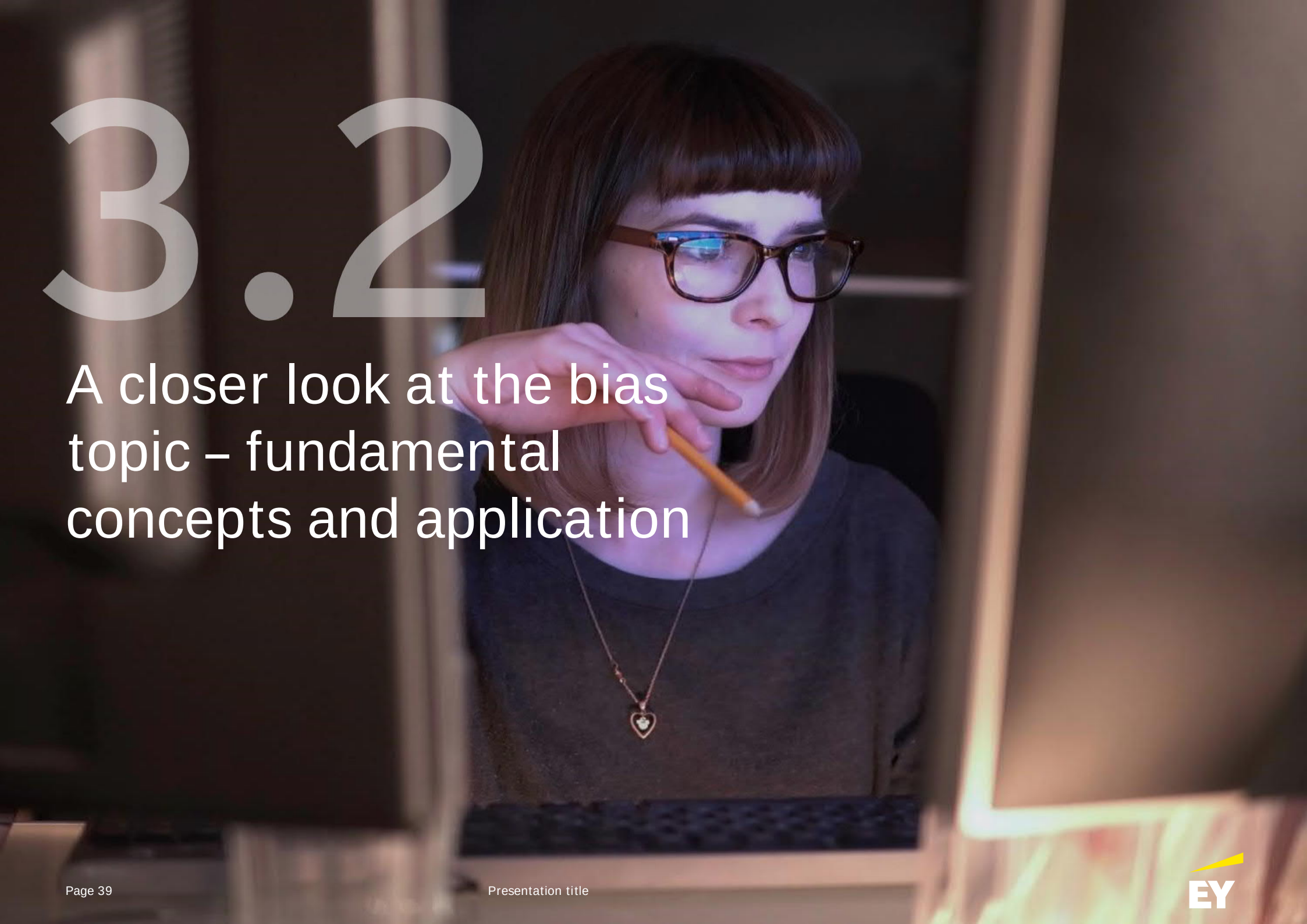
The approach is taken by methods which use shapely values except the fact that shapely values captures feature interactions

Partial Dependence Plots (PDPs)

Shows the impact of one or two variables on predictive outcome at a time. Does not consider feature interactions

Like any software algorithm, LIME/SHAP is not robust against manipulation (e.g. scaffolding*).
In process development, additional controls are therefore required to verify/audit the implementation.

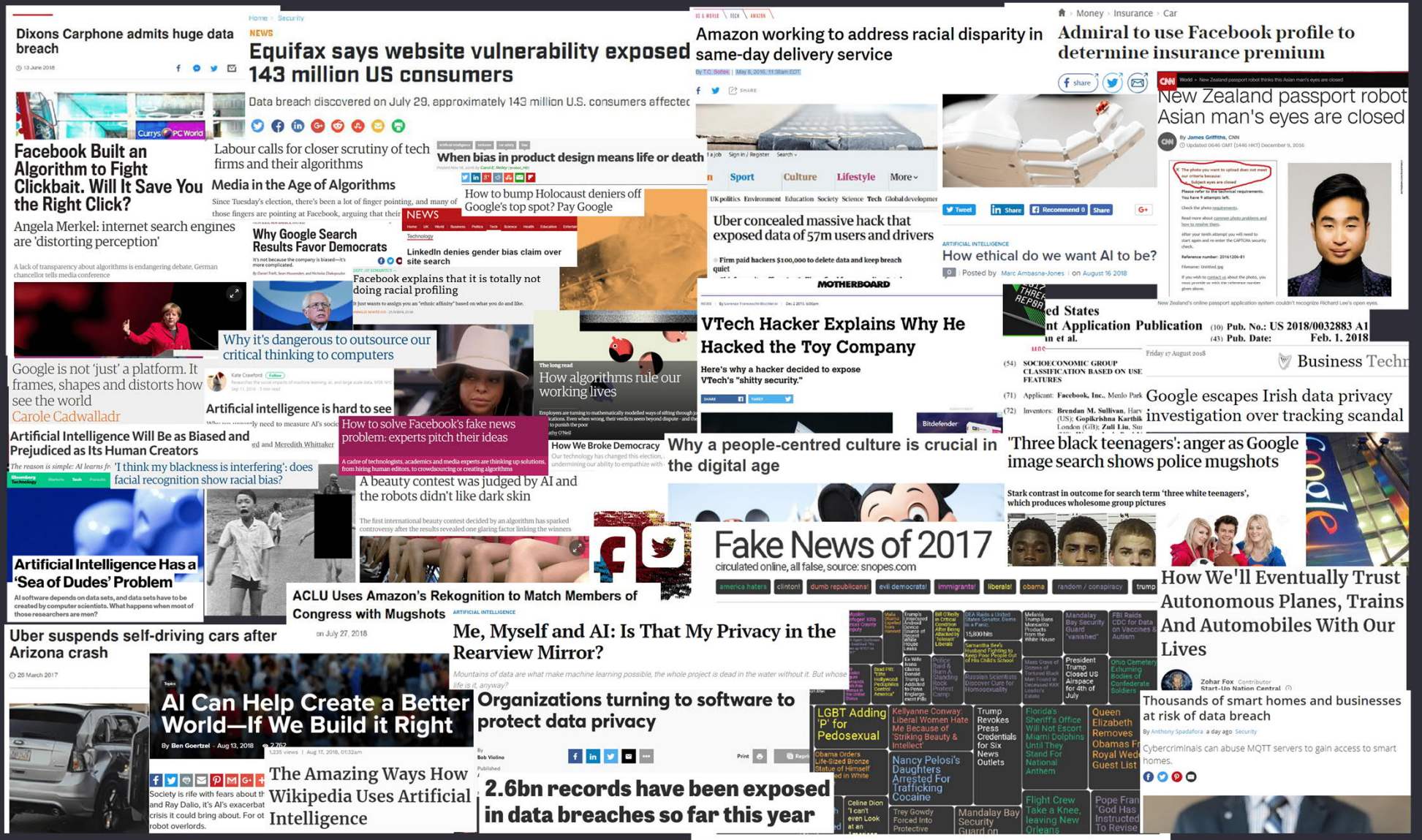
*Source: https://arxiv.org/abs/1911.02508



3.2

A closer look at the bias topic – fundamental concepts and application

Popular perception of AI Ethics: Algorithms in the News



Shaping the policy narrative

UK police use of facial recognition technology a failure, says report

Civil liberties group says systems used by Met and South Wales police are wrong nine times out of 10

▲ One of the floats at last year's
98% of the time. Photograph: T

Police attempts to use

UK police need to slow down with face recognition, says data watchdog

TECHNOLOGY 31 October 2019

By New Scientist staff and Press Association

UK police's facial recognition system has an 81 percent error rate

But officials inside the Metropolitan Police say otherwise.

Rachel England, @rachel_england
07.04.19 in Security

21 Comments

2309 Shares

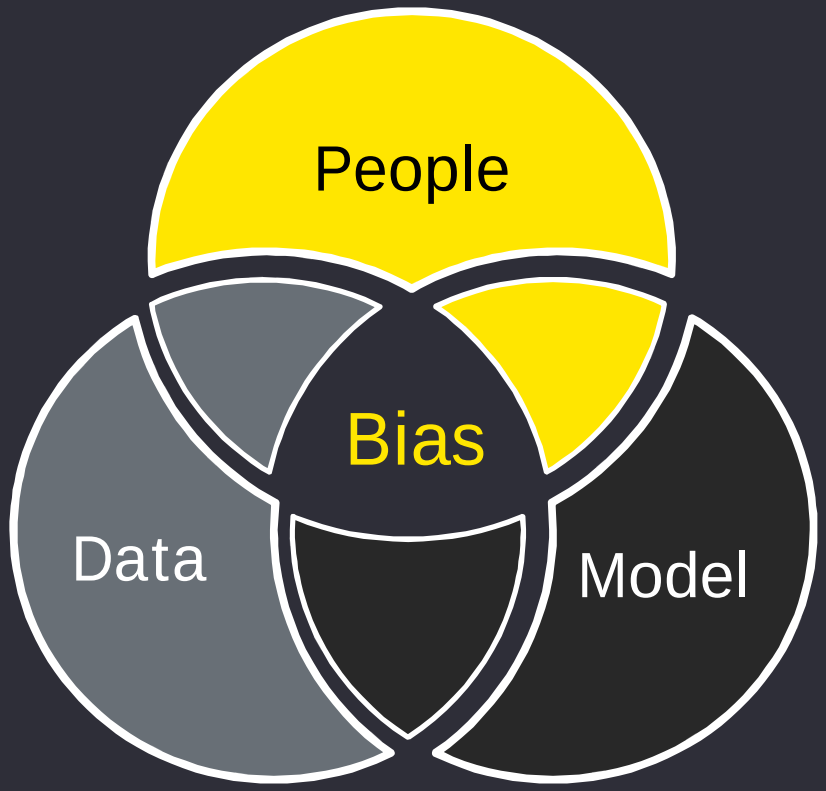
Life
som
Discover
Finance
Stay info
yaho

Sponsore

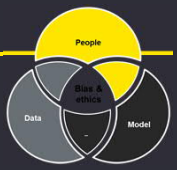
NISSAN

3 core dimensions of bias

Bias can be introduced throughout the AI lifecycle. The following section will focus on bias in data, models and people

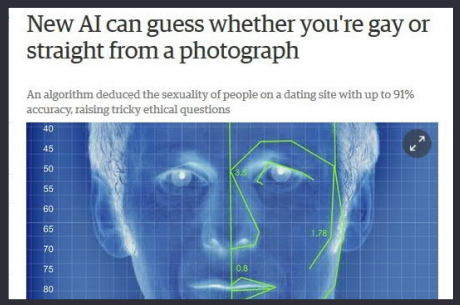


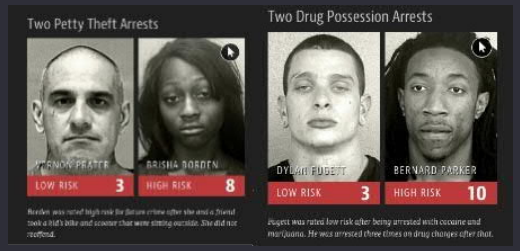
There are different types of biases …

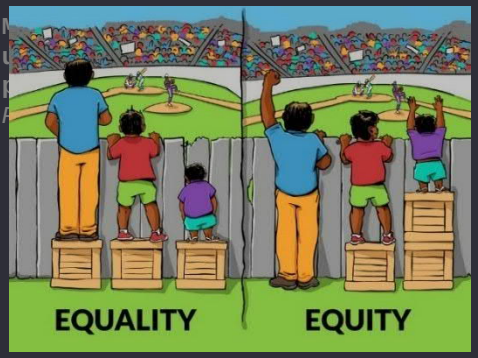


Data

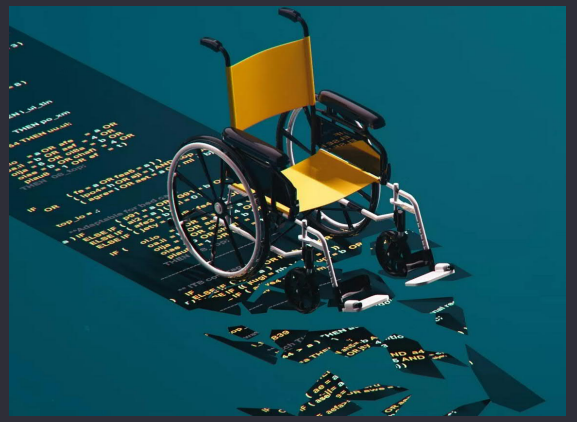
Model







People



Ethics, Fundamental Rights and Societal Impact

All non-trivial* decisions are biased

We seek to minimize bias that is:

- Unintended
- Unjustified
- Unacceptable

as defined by the context where the system is used.

*Non-trivial means the decision space has more than one possible outcome and the choice is not uniformly random.

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Joy Buolamwini & Timnit Begru

Proceedings of Machine Learning Research 81:1-15, 2018

Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

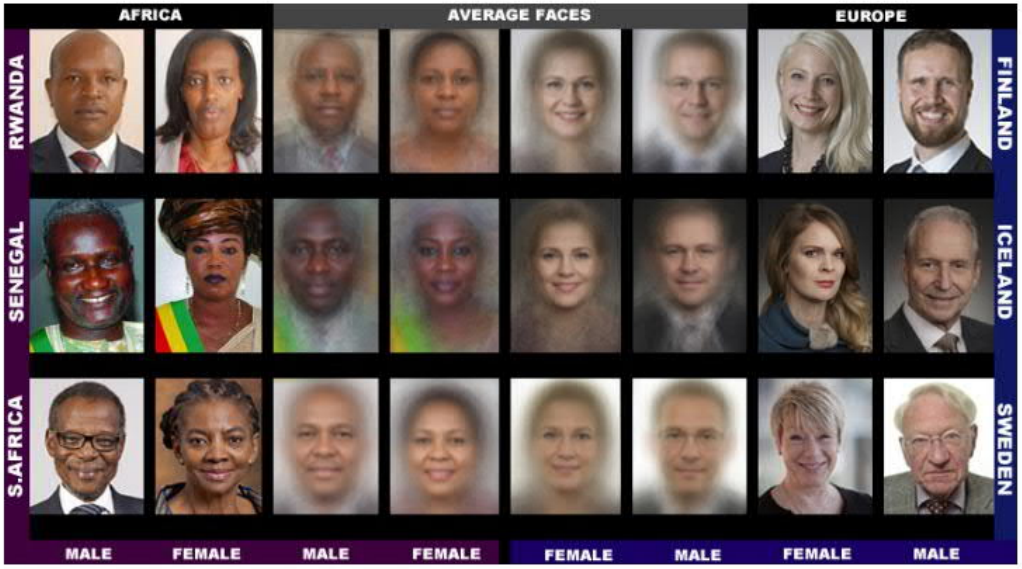
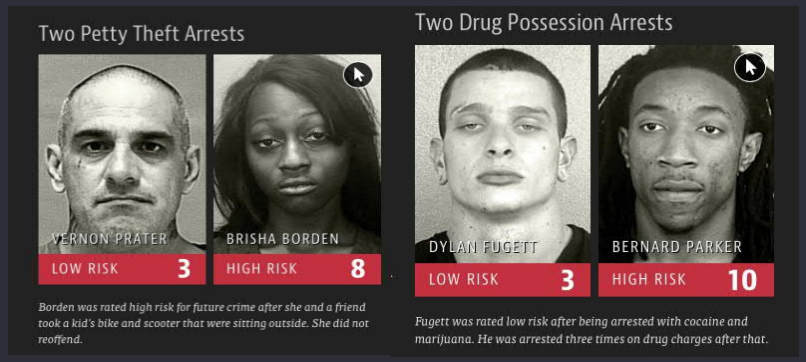
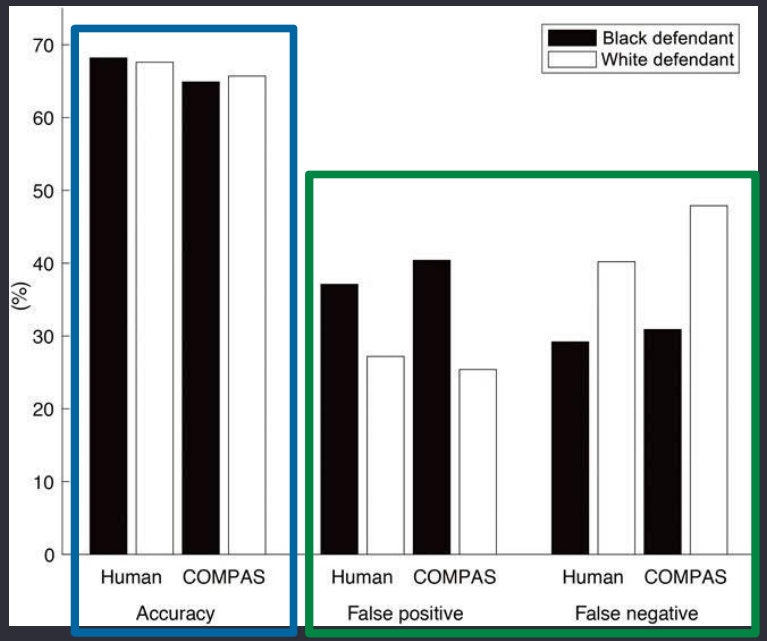


Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

Understanding exactly what bias to avoid also helps to set Algorithmic optimization targets and can lead to more targeted models



Machine Bias: There's software used across the country to predict future criminals. *Propublica*



		Predicted Label: $\hat{y}=1$	Predicted Label: $\hat{y}=-1$	
True Label	$y=1$	True positive	False negative	$P(\hat{y}\neq y \mid y=1)$ False Negative Rate
True Label	$y=-1$	False positive	True negative	$P(\hat{y}\neq y \mid y=-1)$ False Positive Rate
		$P(\hat{y}\neq y \mid \hat{y}=1)$ False Discovery Rate	$P(\hat{y}\neq y \mid \hat{y}=-1)$ False Omission Rate	$P(\hat{y}\neq y)$ Overall Misclass. Rate

When base rates differ, no non-trivial solution can achieve similar FPR, FNR, FDR, FOR.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Dangerously, machine learning is capable of confirming simplistic views of the world: Complex individuals reduced to simplistic binary stereotypes

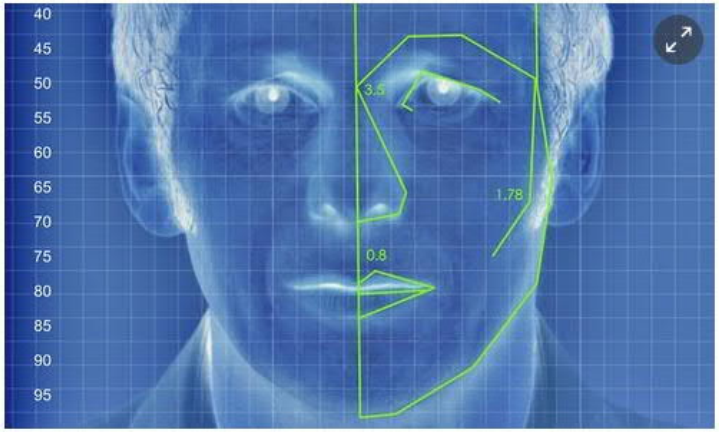
Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.

Contributors: Michal Kosinski, Yilun Wang

Date created: 2017-02-15 04:37 PM | Last Updated: 2017-09-24 04:40 AM

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions



An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

THE CONVERSATION

Academic rigour, journalistic flair

Arts + Culture | Business + Economy | Cities | Education | Environment + Energy | Health + Medicine | Politics + Society | Science + Technology | Brexit

Machine gaydar: AI is reinforcing stereotypes that liberal societies are trying to get rid of

September 13, 2017 11.35am BST

Artem Oleshko/Shutterstock

Email | Twitter 35 | Facebook 136 | LinkedIn 12

Following the old saying that "knowledge is power", companies are seeking to infer increasingly intimate properties about their customers as a way to gain an edge over their competitors. The growth of Artificial Intelligence (AI), algorithms that use machine learning to analyse large multifaceted data sets, provides an especially attractive way to do this. In

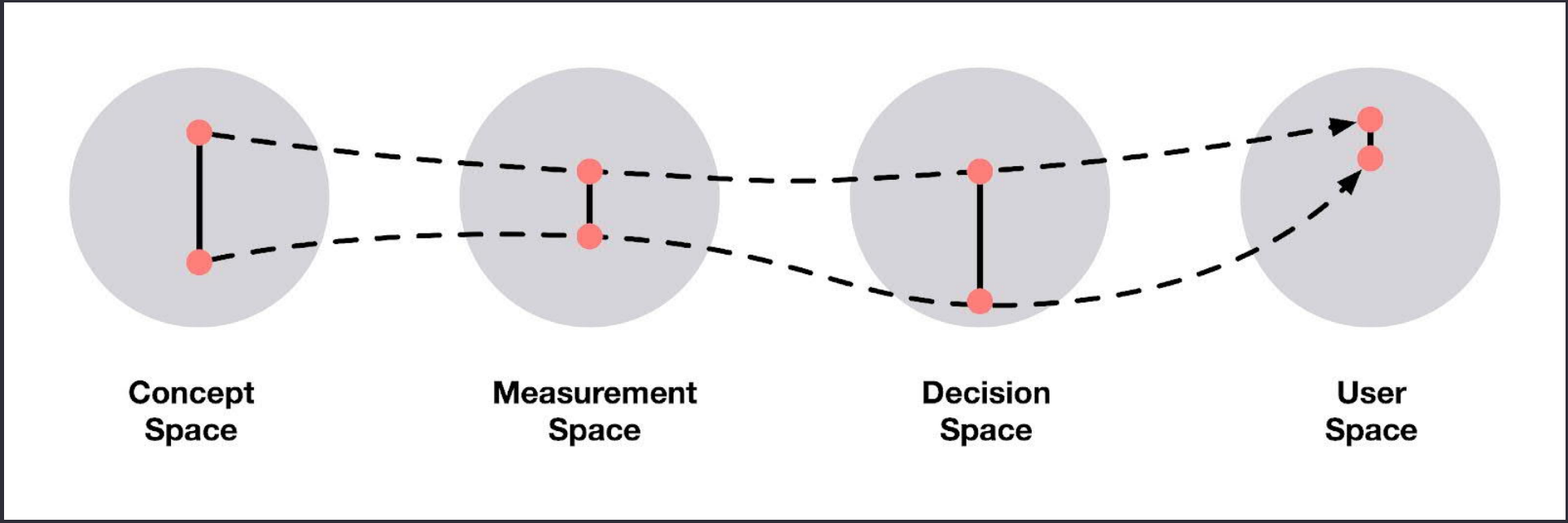
Author

Ansgar Koene
Senior Research Fellow, Horizon Digital Economy, UnBias, University of Nottingham

Bias arising from imperfect mapping between Concept and Execution

- Input data are measurable approximations of conceptually desired factors e.g. impossibility of directly measuring intent of behaviour

- Decision outputs generated by the algorithmic system are interpreted by users, injecting subjective interpretation bias that is unpredictable at the individual level



Key question when developing or deploying an algorithmic system

- Who will be affected?
- What are the decision/optimization criteria?
- How are these criteria justified?
- Are these justifications acceptable in the context where the system is used?

So we agree that we all want fair lending?

What would you like to achieve (gender as an example)

- For every 100 men who are granted a loan, 100 women should also be granted a loan (would this definition work also for a racial bias / nationality bias setting?)
- If 57% of all male applicants are granted a loan, then at least also 57% of all female applicants should be granted a loan (just in case most families left loan organization to females … would that mean that among single men/women the acceptance rate is then biased anyway?)
- Among men and women with the same risk characteristics (i.e. model feature values), the same percentage should be granted a loan (what do you do in case societal sexism precludes female applicants from getting to the same e.g. salary levels, hence it is much harder for females to get to a certain level of model feature values for some features?)

There are several different (mathematical, precise) definitions of fairness, into which the above ideas could be translated. It can be proven that these notions will conflict in certain situations. I.o.w., you must make a choice for one fairness concept and then you remain vulnerable to accusations of not being fair against the others

… and how do you make sure you're fair, or even defend in public?

So you have chosen a fairness definition to control nationality or race bias, that you want to pursue for your credit application decision algorithm.

- Neither race nor nationality are typically polled, stored or used for credit risk analysis. E.g. the German Credit Act (Kreditwesengesetz) forbids this
- Would you be able to measure potential indirect bias, i.e. to detect race/nationality bias without using race/nationality data directly, e.g. through publicly available demographic/statistical information on correlations between race/nationality and features represented in your model?
- What if the data used to train your model already contains bias
- Or would you rather try to use algorithms susceptible to indirect bias in areas where the public just isn't that interested, or where they are potentially less relevant indeed?

On Bias: Some relevant conceptual and academic observations

A. “Many analysts, one dataset: Making transparent how variations in analytical choices affect results”; Silberzahn, Uhlmann, Nosek et (many) all

- The outcome is influenced by the choice of ML/modelling approach
- The outcome is influenced by the features used in each approach
- Aggregating different approaches may provide the certainty one is looking for
- Existence of a bias effect does not imply insight in its causes

B. “Delayed Impact of Fair Machine Learning”; L.T. Liu, S. Dean, E. Rolf, M. Simchowitz and M. Hardt (Berkeley Artificial Intelligence Research - BAIR)

- Feedback loops may actually make that well intended policies are in fact creating negative consequences at the aggregate/society level
- Such societal costs/benefits could provide an alternative objective function in model calibration

C. “What do you do when data tells you to be racist?” ; a blogpost on towardsdatascience.com by Matthew Stewart (and papers it is based on)

- Different notions of fairness can be expressed in mathematical conditions
- These mathematical conditions can be shown to not be a priori mutually consistent

Many Analysts, one Dataset – research suggests that statistical freedom, personal behaviour and bias conclusions interact, is objectivity an illusion?

Experiment with 29 academic teams, same dataset, one research question:

Is there a bias that leads to dark skinned football players getting a red card more easily than lighter skinned colleagues?

Very different results across teams; but aggregating results clearly reveals bias. Differences in outcome must be due to:

- different analytical approaches
- interpretational differences
- Experience/Prior knowledge
- Prior beliefs maybe (subjectivity)

- On the graph:
  - Up to value of 1 means no effect
  - while >1 mean bias
  - 9 teams found no significant bias
  - while 20 teams did find an effect whereby darker skinned players get a red card more frequently

Team	Analytic Approach	OR
12	Zero–inflated Poisson regression	0.89
17	Bayesian logistic regression	0.96
15	Hierarchical log–linear modeling	1.02
10	Multilevel regression and logistic regression	1.03
18	Hierarchical Bayes model	1.10
31	Logistic regression	1.12
1	Ordinary least squares with robust standard errors, logistic regression	1.18
4	Spearman correlation	1.21
14	Weighted least squares regression with clustered standard errors	1.21
11	Multiple linear regression	1.25
30	Clustered robust binomial logistic regression	1.28
6	Linear Probability Model	1.28
26	Three–level hierarchical generalized linear modeling with Poisson sampling	1.30
3	Multilevel Binomial Logistic Regression using bayesian inference	1.31
23	Mixed model logistic regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear probability model, logistic regression	1.34
5	Generalized linear mixed models	1.38
24	Multilevel logistic regression	1.38
28	Mixed effects logistic regression	1.38
32	Generalized linear models for binary data	1.39
8	Negative binomial regression with a log link analysis	1.39
20	Cross–classified multilevel negative binomial model	1.40
13	Poisson Multi–level modeling	1.41
25	Multilevel logistic binomial regression	1.42
9	Generalized linear mixed effects models with a logit link function	1.48
7	Dirichlet process Bayesian clustering	1.71
21	Tobit regression	2.88
27	Poisson regression	2.93

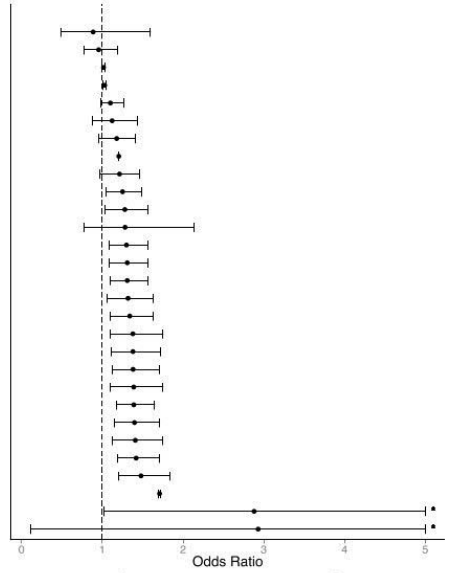


Fig. 1. Point estimates and 95% confidence intervals for analysis teams for the primary research question: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players? Note that the asterisks correspond to a truncated upper bound for Team 21 (11.47) and Team 27 (78.66) to increase the interpretability of this plot.

Democratizing model development? A potential conclusion could be that building challenger models might be better than challenging models

Differences were in algorithm/methodologies, but also use of different covariates / features in teams' analyses. Only one feature, #games played, was used by all

Covariate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	20	21	23	24	25	26	27	28	30	31	32	% used
Position	■	■	■						■	■	■	■	■	■					■	■		■	■	■		■	■	■	■	62%
Height	■	■		■		■																					■	■		38%
Weight	■	■		■							■	■										■	■				■	■		38%
Age	■	■				■					■			■	■				■											24%
League Country	■														■					■					■					17%
Goals		■									■				■									■	■					17%
Referee Country		■	■	■					■	■																		■		17%
Victories		■	■							■														■						10%
Club	■													■																7%
Referee	■													■																7%
Player Cards																	■	■												7%
Player																	■	■												3%
Referee Cards																	■													3%
Draws										■																				3%
N Covariates	7	6	2	3	0	3	0	0	2	3	3	2	1	6	1	2	2	2	1	3	2	3	4	6	1	2	3	4	1	

Table 2. This overview shows the covariates used by each team. Team numbers are listed on the top and covariates on the left. A shaded box indicates that the corresponding team used the covariate in their final model. The table is ordered by the frequency by which each covariate was used.

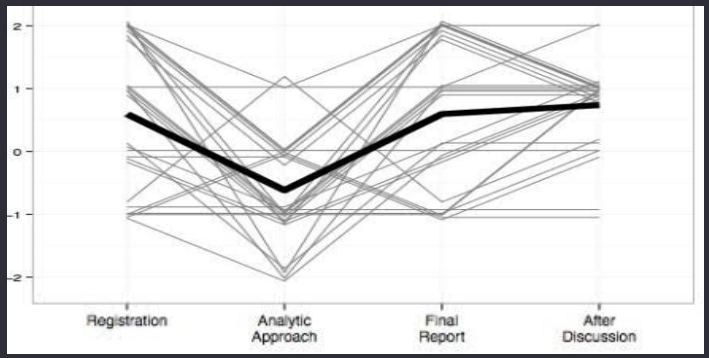
Two variables which were securing initially identification purposes, were used as covariates by some teams; even after their potentially distorting impact was discussed (League & Club at the moment the data were collected)

One team reported that removing 7 outliers (0.3 % of the dataset) made the effect go away. Other teams subsequently checked and found no impact

- Always do an impact assessment when removing outliers (or remediating data issues for that matter)

Subjective beliefs of teams were tracked (impacted by interactions and results, but also shaping research)

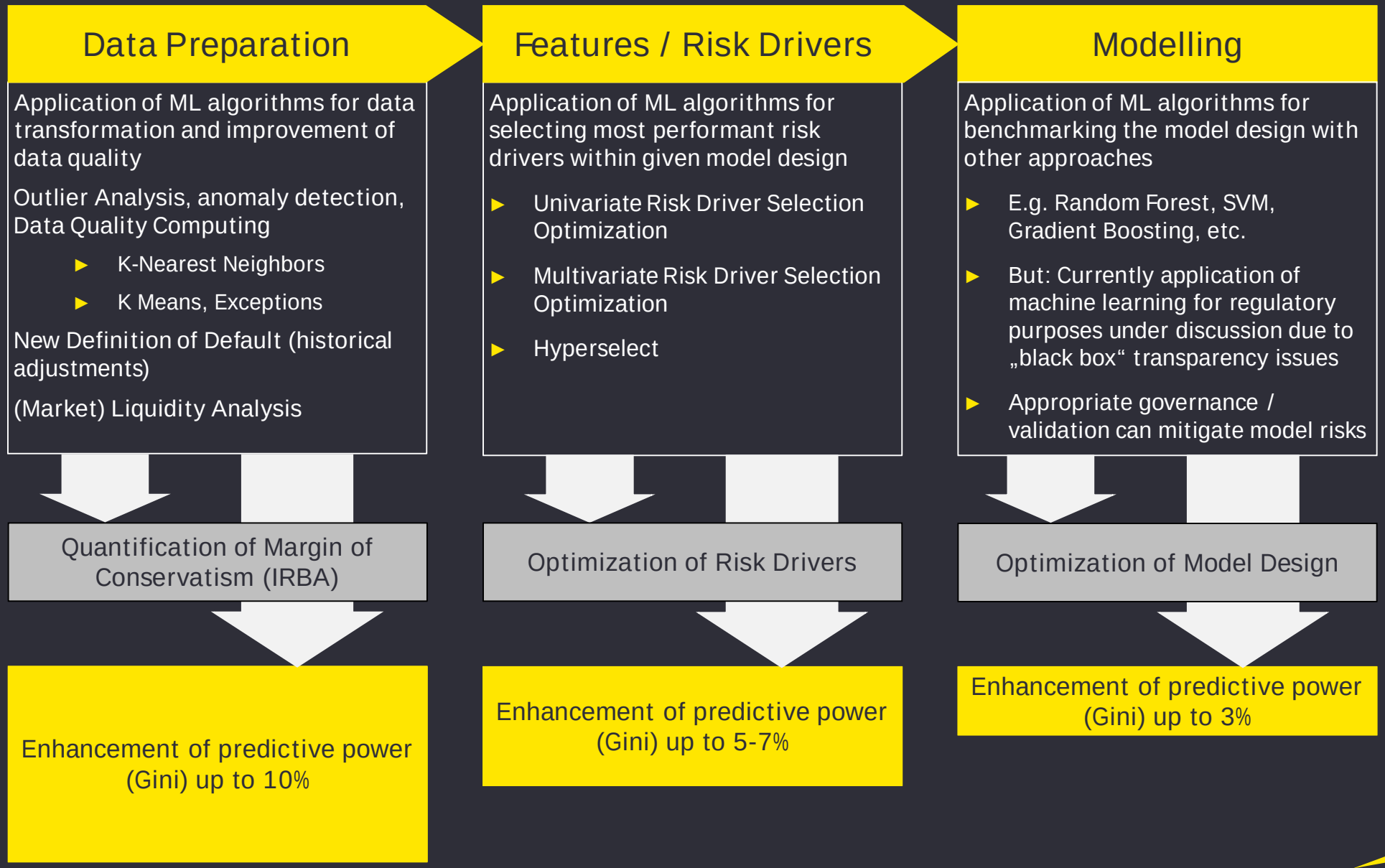
- Most teams agreed their analysis did not show causality
- Initial beliefs only weakly correlated to final results

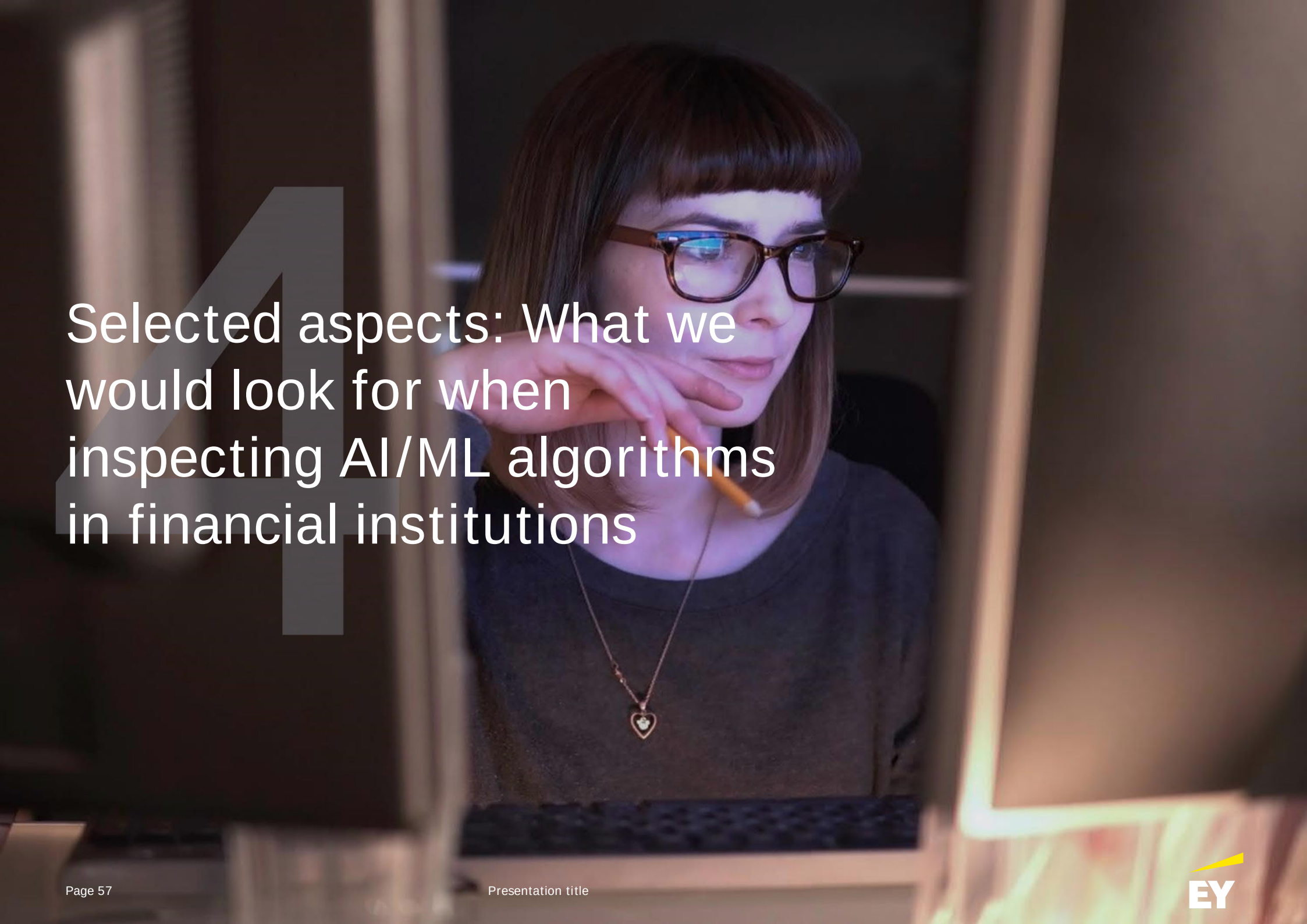


3.3

Conclusions, if any?

How far down the assembly line should you go, if most of the trouble can be avoided by leaving out the last step?

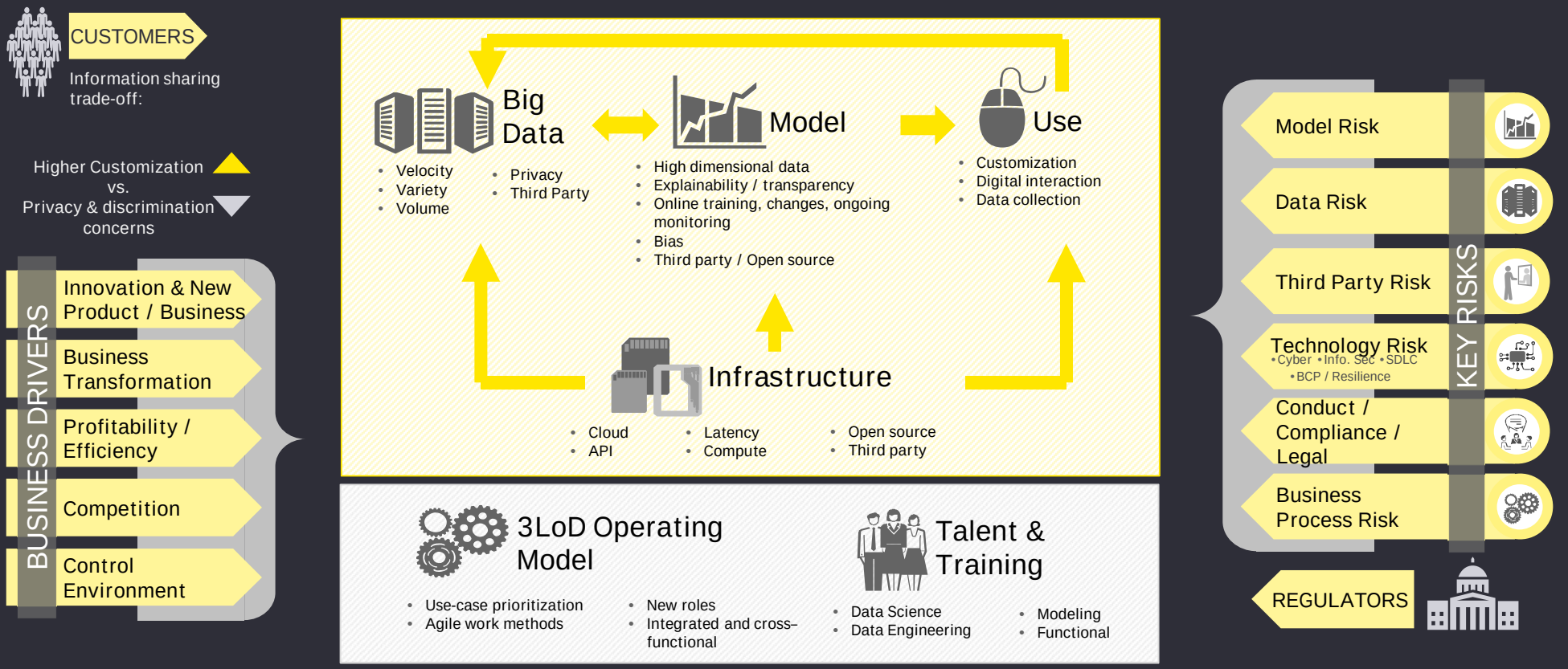




4 Selected aspects: What we would look for when inspecting AI/ML algorithms in financial institutions

AI/ML ecosystem – key characteristics and risks

Dynamic, connected, interactive nature of AI ecosystem poses risks, with widespread impact



Auditing ML – areas of perceived higher risk

The following topics seem to give rise to higher risks for ML/AI than for traditional models/algorithms. This has no effect on the additional classical model inspection topics.

1. Specific data and vulnerability topics (new types of data and IT implementation forms potentially in use)
2. Governance including ethics (recent experimental phase)
3. Explainability (nature of algorithms)
4. Bias (not really different for traditional models, but less easily diagnosed maybe, hence connected with model transparency/explainability)
5. Self-Learning properties (a model change topic, which relates back to Governance)
6. Hidden from view (big-tech in the value chain, outside directly supervised realm in some way, or indirect in their influence)

Data governance and data quality – is it really applied to all data that are needed to build, validate and maintain ML/AI models, …

Domain	Component	Level1: Initial	Level2: Repeatable	Level3: Defined	Level4: Managed	Level5: Optimizing
Organisational Model	Organisation design	There is no/limited dedicated data governance organisation. There is lack of awareness and ownership of data quality issues.	There is some awareness of data quality issues. Resources are allocated and organisational structure emerges.	Management buy-in exists. Business is actively engaged with technology with defined stewardship and ownership roles and responsibilities.	Executive sponsorship exists. Enterprise data governance structure is established. Data is a part of defined roles. Personnel are measured against DQ.	Defined executive level responsibilities over the quality of data exists. Data governance is continuously measured and monitored. ROI is tracked.
	Data Governance Roles and Responsibilities	No or few roles are established on data ownership. Data as a result is inconsistent.	Fragmented roles exist, yet with no clear definition or mandate.	Data Ownership defined by business SMEs (mid-management) with technical support for critical data elements.	Data Ownership managed by business SMEs in addition to partially established data stewardship structure.	Data Owners and Data Stewards with clearly defined roles and responsibilities form a robust data governance committee.
	Communication	Lack of communication is prevalent across the enterprise.	Siloed but repeatable communication processes exist with a subset of data related initiatives. In most cases communication is technology focused with minimal business input.	Data related decisions are made with adequate level of representation across the enterprise, yet conflict resolution processes is not fully established.	Data related decisions are made with the appropriate level of representation from across the firm with clear processes and procedures for conflict resolution and communication.	Communication/awareness plan for informing the business, operations and IT communities about the data governance efforts and the impact and benefit to supported initiatives is established and implemented.

… and are there clear standards and policies for data that are not in the classical sense “native” to the institute (social media, web, unstructured)

Domain	Component	Level1: Initial	Level2: Repeatable	Level3: Defined	Level4: Managed	Level5: Optimizing
Policies and Standards	Data Policies	Policies are not clearly defined across the firm, leading to inconsistent data.	Fragmented policies exist in silos, yet are not established across the Enterprise and are repeatable.	Policies are defined by the enterprise for key focus areas (Element Based Management, Metadata, Reference Data, Data Quality, SLAs, Data Ownership and Data Issue Management).	Policies are defined by the enterprise for key focus areas and a data governance structure is in place although not fully active in management and enforcement.	Policies are defined and established with the data governance committee actively managing the enforcement across the Enterprise.
	Data Standards	There are no clearly defined and established standards.	Some standard processes exist, but they are applied on an ad-hoc basis. Taxonomy and metadata are developed and managed for critical data elements.	There are continued initiatives to enrich taxonomy and metadata to standardize data access, usage, and accountability.	Data definitions and business rules are maintained as part of enterprise information management. Data governance is embedded in corporate project methodologies.	Standards are defined and established with the data governance committee actively managing the enforcement across the Enterprise
Processes and Procedures	Processes and Procedures	There is a lack of defined processes to address data quality.	Data governance is reactive and data-related controls and processes are project focused. Remediation workflows are repeatable but not clearly defined.	Data governance is used to establish consistency in the organisation's approach to data quality and information management.	Repeatable and consistent processes exist. Data governance processes are incorporated in enterprise change management initiatives and managed.	Data governance processes are automated with procedures clearly defined. Processes are reviewed and updated to reflect changing business objectives. Processes are mapped to Data Governance roles (RACI matrices).

AI and Data Protection/Privacy regulations

GDPR

GDPR central principles apply;

- Legal basis
- Transparency
- Privacy by design
- Etc…

Article 29 Data Protection Working Party Guidelines – Focus on big data, machine learning and AI

Article 4.11 ”Profiling” and Article 22 Decision based solely on “automated processing”, including “profiling”

- General prohibition
- Certain exceptions, e.g. express consent

New E-Privacy Regulation

Current E-Privacy and Communications Directive (implemented in Ecom Act) targets telecommunications operators

New E-Privacy Regulation – was planned to come into effect during 2019 but will be delayed

Does not only apply to providers of electronic communications services but also for example;

- Website owners
- Owners of apps that have electronic communication as a component
- Natural or legal persons sending direct marketing communications (incl new rules regarding use of cookies)

AI and Security regulations

AI Security

- Security and cybersecurity threats have in general increased significantly the last few years
- AI is used for cybersecurity attacks, terrorism etc
- AI is used to protect businesses against cybersecurity attacks, terrorism etc
- Using AI for business purposes opens up for vulnerability

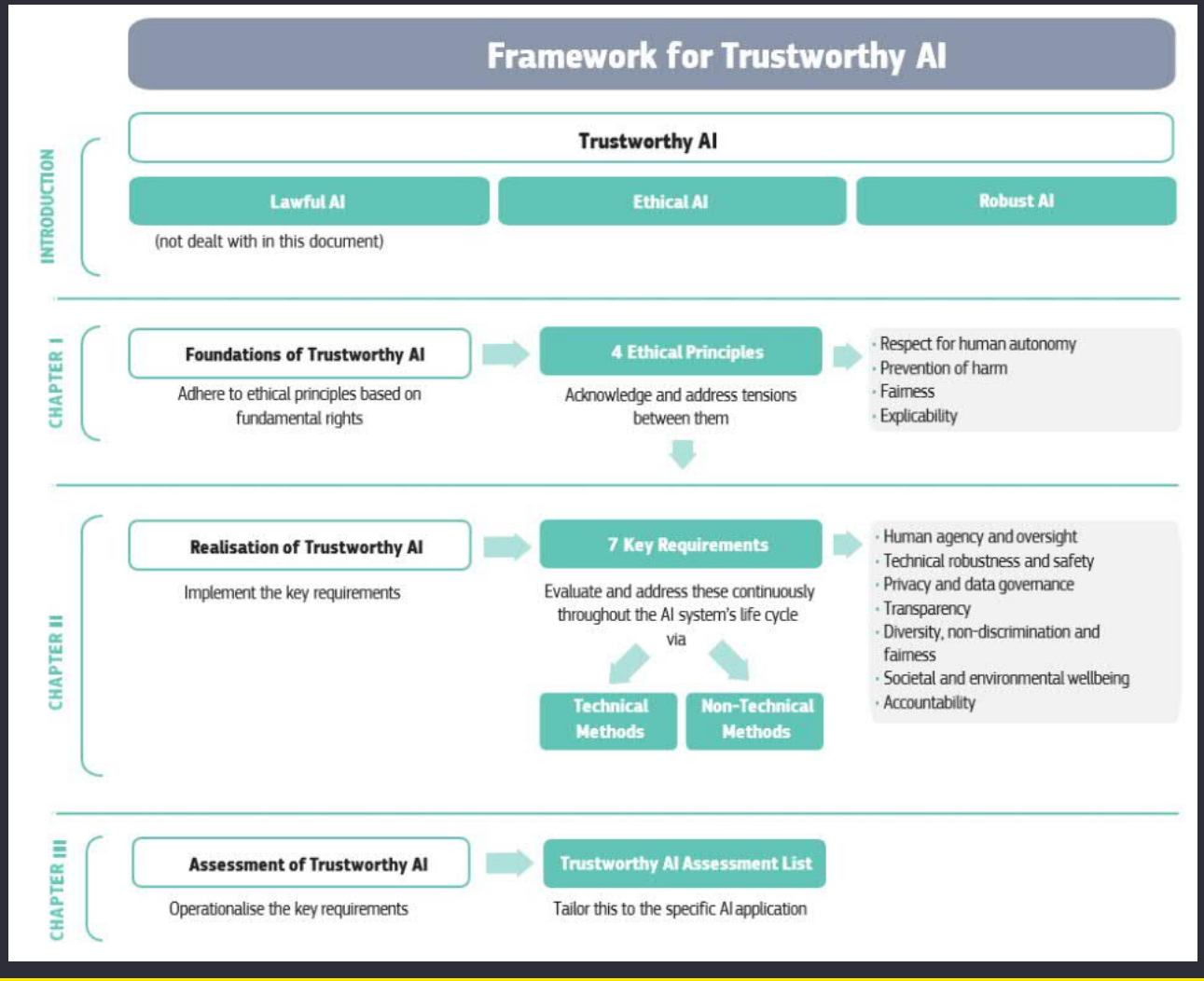
→Compliance with security regulations and general robustness is crucial for trustworthy AI within a company

Security regulations

- NIS Directive and Information Security Act (valid from 1 Aug 2018)
- New Security Act (April 2019) – targeted at “national security”, protection against terrorism etc.
- Regulations targeted for ”national defense” – may indirectly be relevant to several sectors

Governance and Ethics – ethical considerations appropriately recognized and reflected with reference to relevant guidelines?

e.g. the EU “Ethics Guidelines for Trustworthy AI”



Institutions should be able to position themselves against the 7 key requirements deemed relevant for Trustworthy AI

Is AI/ML integrated into normal Model Risk Management across all of the model lifecycle, including model definition, inventory, etc … ?

What are the considerations to accelerate and scale the use of AI/ML in a controlled manner?

1

- Business Strategy
- Risk Appetite
- AI/ML Adoption Strategy

2 AI/ML Definition and Identification

- Establish AI/ML definition to ensure consistent population capture
- Leverage and enhance existing model inventory management process to identify AI/ML applications; link with:
  - Innovation programs
  - New product approval (NPA) / business activity
  - Third party sourcing
  - IT/Tech operations
  - ML technique users
  - EUC program

3 Enhance Existing Risk and Control Frameworks

- Enhance existing risk and control frameworks
- Consider cross-functional governance and controls
- Roll-out enterprise wide training and communication
- Risk based application of the risk and control frameworks
  - New Product Approval (NPA) like process for high risk cases
  - Fast-track process for low risk cases

4 Operating Model (Interim State to Accelerate and Scale)

- Leverage existing governance framework and committees for 2nd line oversight
- Consider AI COE and AI Champions (1st line) based on existing innovation functions and model development units
- Establish Board and senior management reporting on AI/ML strategy, implementation, governance and controls

5 Capabilities

- Develop projections for AI/ML adoption for capability assessments and prioritization
- Enhance capabilities for AI/ML modeling and implementation:
  - Data management
  - Infrastructure
  - Training and hiring strategy

Note: The above components of the AI/ML governance are not sequential and some aspects will be in parallel to establish and implement the framework

A resilient AI/ML ecosystem, will accelerate the adoption of AI/ML by establishing stakeholder trust and confidence

Areas to tackle

- Is model materiality spelled out in a useful manner also for AI/ML?
- Is the institute clear about minimum standards concerning model explainability and interpretability, to tackle the topics from section 3.1
- Has the institute reflected its approach to fairness/bias where applicable, to tackle the topics from section 3.2
- Are all ML/AI models clearly subject to the model change policies that might be relevant for their area of application
- Is the bank capturing all (hidden?) ML/AI algorithms that effectively impact their decision making and/or predicions, in line with its governance ambitions

Disruptors of the Banking sector?

Avoid banking sector regulations by being classified at tech-platform

- Copy examples set in other sectors by Uber, Airbnb, etc.

Fintech startups willing to take risks with using AI

- No reputation that needs protecting

Big-tech forays into financial services, building on access to data and AI infrastructure

- Libra (Facebook)
- Payment services (Apple Pay, Google Wallet, Alipay)
- Providing of small business loans (Amazon, Alibaba)

“

Is it enough to supervise banks in order to supervise banking?

EY | Assurance | Tax | Transactions | Advisory

About the global EY organization
The global EY organization is a leader in assurance, tax, transaction and advisory services. We leverage our experience, knowledge and services to help build trust and confidence in the capital markets and in economies the world over. We are ideally equipped for this task – with well trained employees, strong teams, excellent services and outstanding client relations. Our global purpose is to drive progress and make a difference by building a better working world – for our people, for our clients and for our communities.

The global EY organization refers to all member firms of Ernst & Young Global Limited (EYG). Each EYG member firm is a separate legal entity and has no liability for another such entity's acts or omissions. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. For more information about our organization, please visit ey.com.

In Germany, EY has 20 locations. In this publication, "EY" and "we" refer to all German member firms of Ernst & Young Global Limited.

BHEIN 20-02-06
ED None

ey.com/de