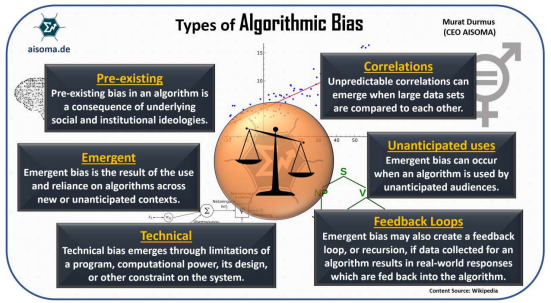
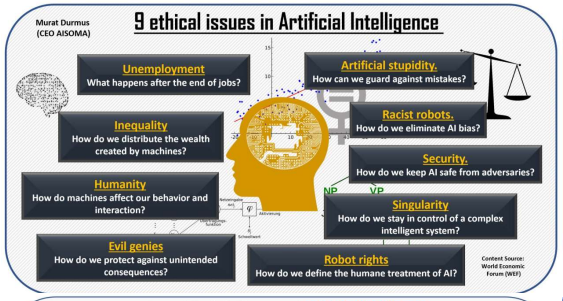
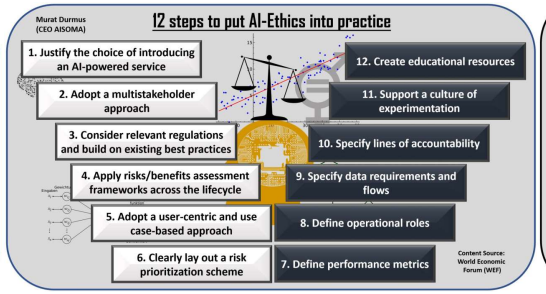
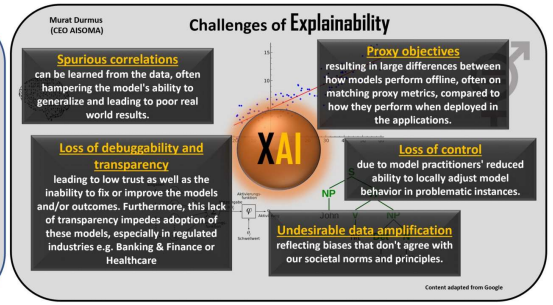
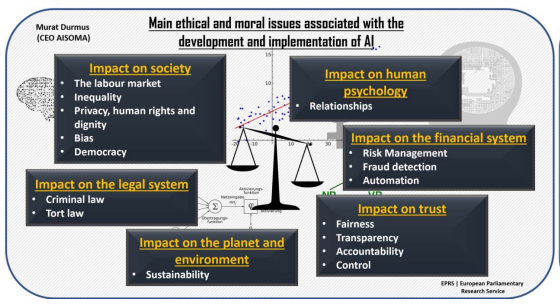
13 Useful Slides on AI, AI-Ethics, & XAI

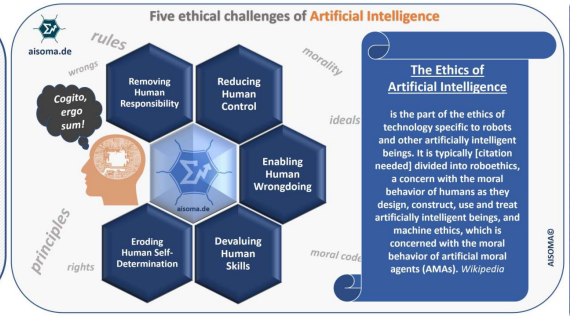


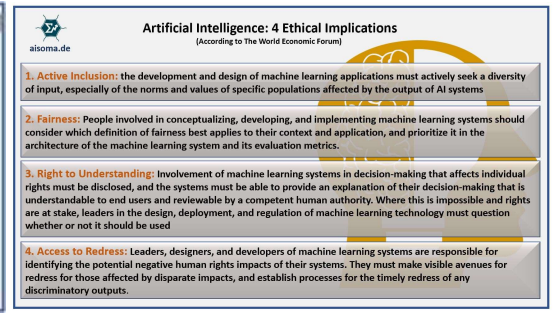






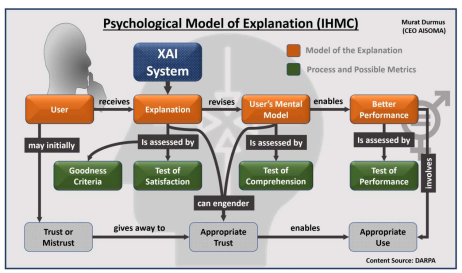


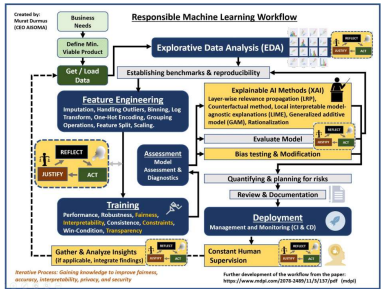


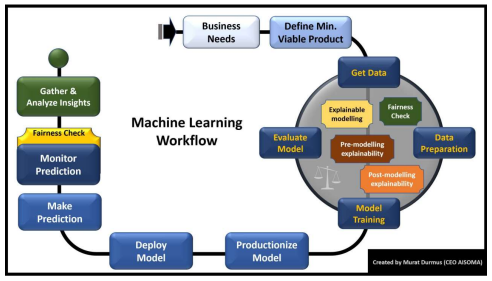


An overview of model agnostic, local, and post hoc Interpretability Methods
Murat Durmus (CEO AISOMA)

Method	Description	Pros	Cons
LIME	LIME, or Local Interpretable Model-Agnostic Explanations, is an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.	- Fully model-agnostic - Many contributors	- The inclusion of unrealistic data instances - Ambiguity of how to select kernel width - Coverage not always clear
SHAP	SHAP, or SHapley Additive exPlanations, is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.	-It can be seen how both functions move predictions up & down - Based on solid theory. SHAP values have useful and proven mathematical properties - The TreeExplainer is fast	- The inclusion of unrealistic data instances - Computationally expensive in the general case - Based on solid theory. It can be hard to dive into all the math inside this method
Permutation Importance	Permutation Importance is an intuitive way to assess the impact of a feature on the black-box model performance.	- Simple and intuitive - Available through the eli5 library (Python) - Easy to compute	- Labeled test data to compute the loss are required - Different shuffles may give different results - Greatly influenced by correlated features
Partial Dependency Plot	One of the simplest and most understandable methods is a partial dependency graph. This dependency graph shows us the dependency of a target variable from a particular feature.	- Easy and intuitive - Available in sklearn (Python)	- Assumption of feature independence - Loss of higher order interactions
Anchor	A novel model-agnostic system that explained the behavior of complex models with high-precision rules called anchors, representing local, "sufficient" conditions for predictions.	- Clear coverage - High precision - The anchors are expressive	- The inclusion of unrealistic data instances - The code available in the original repository is still in progress







Created by Murat Durmus (CEO AISOMA)

Types of Algorithmic Bias

Murat Durmus
(CEO AISOMA)

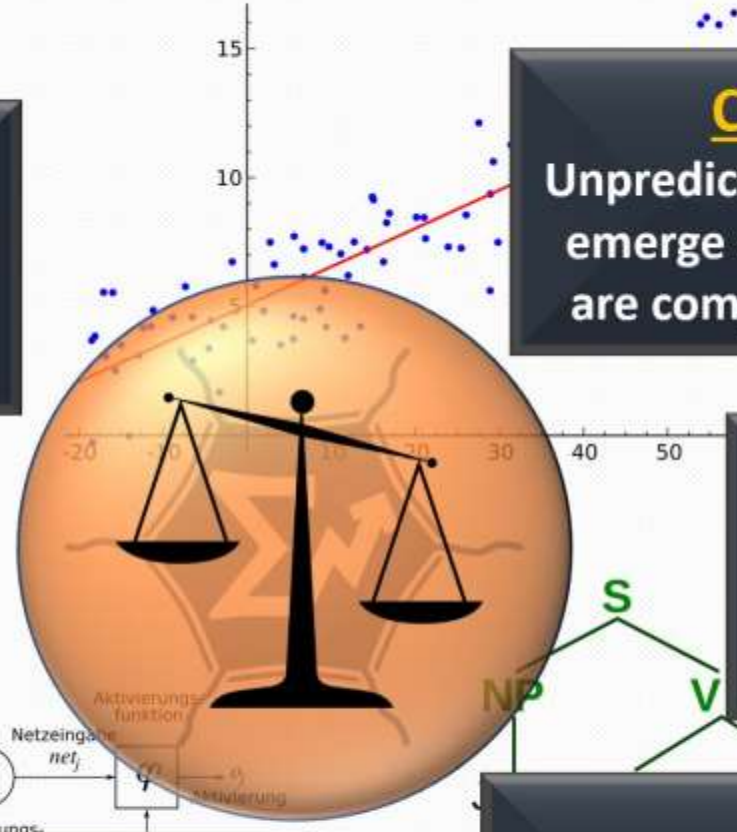
aisoma.de

Pre-existing

Pre-existing bias in an algorithm is a consequence of underlying social and institutional ideologies.

Correlations

Unpredictable correlations can emerge when large data sets are compared to each other.



Emergent

Emergent bias is the result of the use and reliance on algorithms across new or unanticipated contexts.

Unanticipated uses

Emergent bias can occur when an algorithm is used by unanticipated audiences.

Technical

Technical bias emerges through limitations of a program, computational power, its design, or other constraint on the system.

Feedback Loops

Emergent bias may also create a feedback loop, or recursion, if data collected for an algorithm results in real-world responses which are fed back into the algorithm.

Content Source: Wikipedia

Murat Durmus
(CEO AISOMA)

9 ethical issues in Artificial Intelligence

Unemployment
What happens after the end of jobs?

Inequality
How do we distribute the wealth created by machines?

Humanity
How do machines affect our behavior and interaction?

Evil genies
How do we protect against unintended consequences?

Artificial stupidity.
How can we guard against mistakes?

Racist robots.
How do we eliminate AI bias?

Security.
How do we keep AI safe from adversaries?

NP VP

Singularity
How do we stay in control of a complex intelligent system?

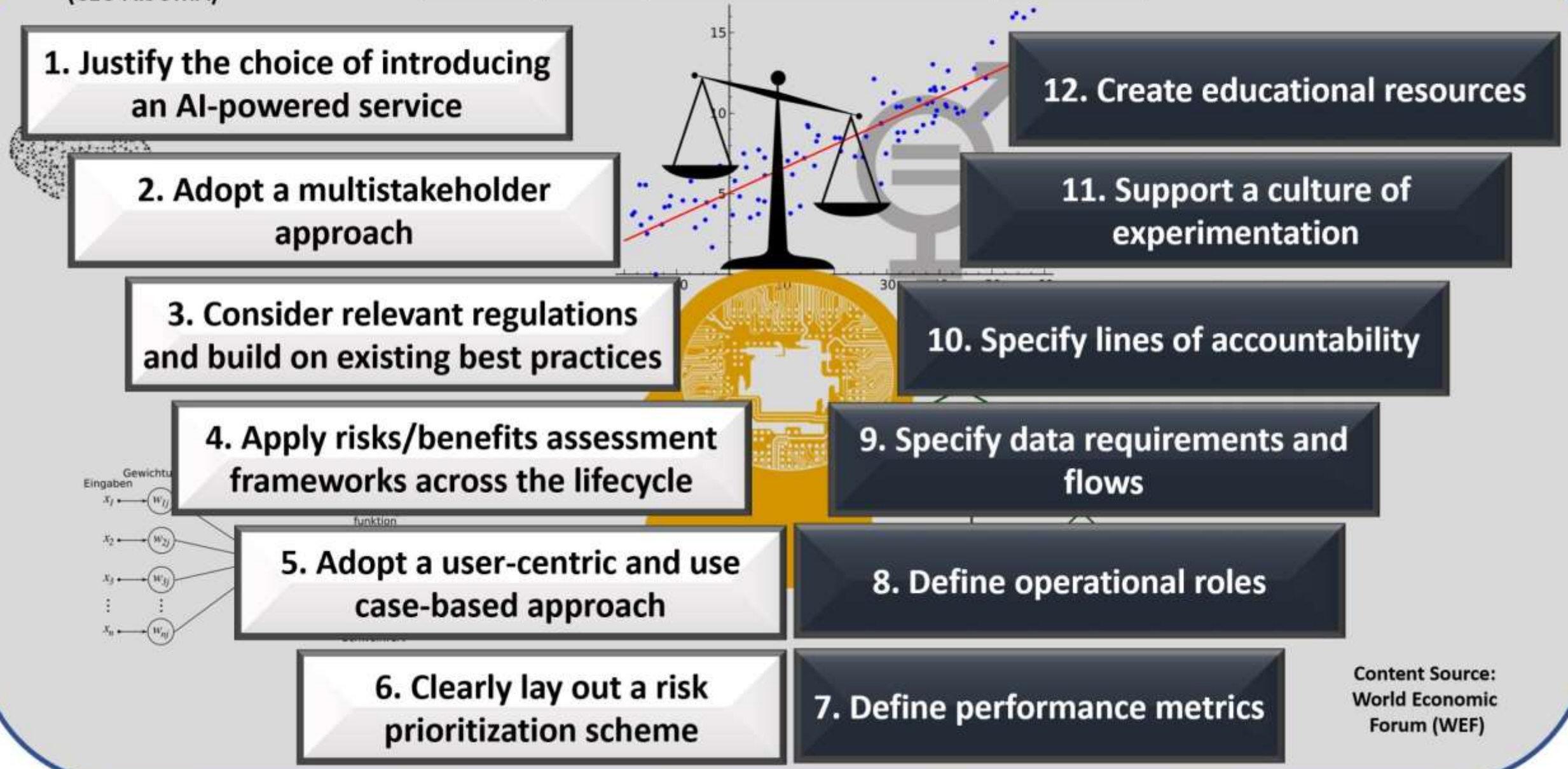
Robot rights
How do we define the humane treatment of AI?

Content Source:
World Economic
Forum (WEF)

Murat Durmus
(CEO AISOMA)

12 steps to put AI-Ethics into practice

1. Justify the choice of introducing an AI-powered service
2. Adopt a multistakeholder approach
3. Consider relevant regulations and build on existing best practices
4. Apply risks/benefits assessment frameworks across the lifecycle
5. Adopt a user-centric and use case-based approach
6. Clearly lay out a risk prioritization scheme
7. Define performance metrics
8. Define operational roles
9. Specify data requirements and flows
10. Specify lines of accountability
11. Support a culture of experimentation
12. Create educational resources



Content Source:
World Economic Forum (WEF)

Murat Durmus
(CEO AISOMA)

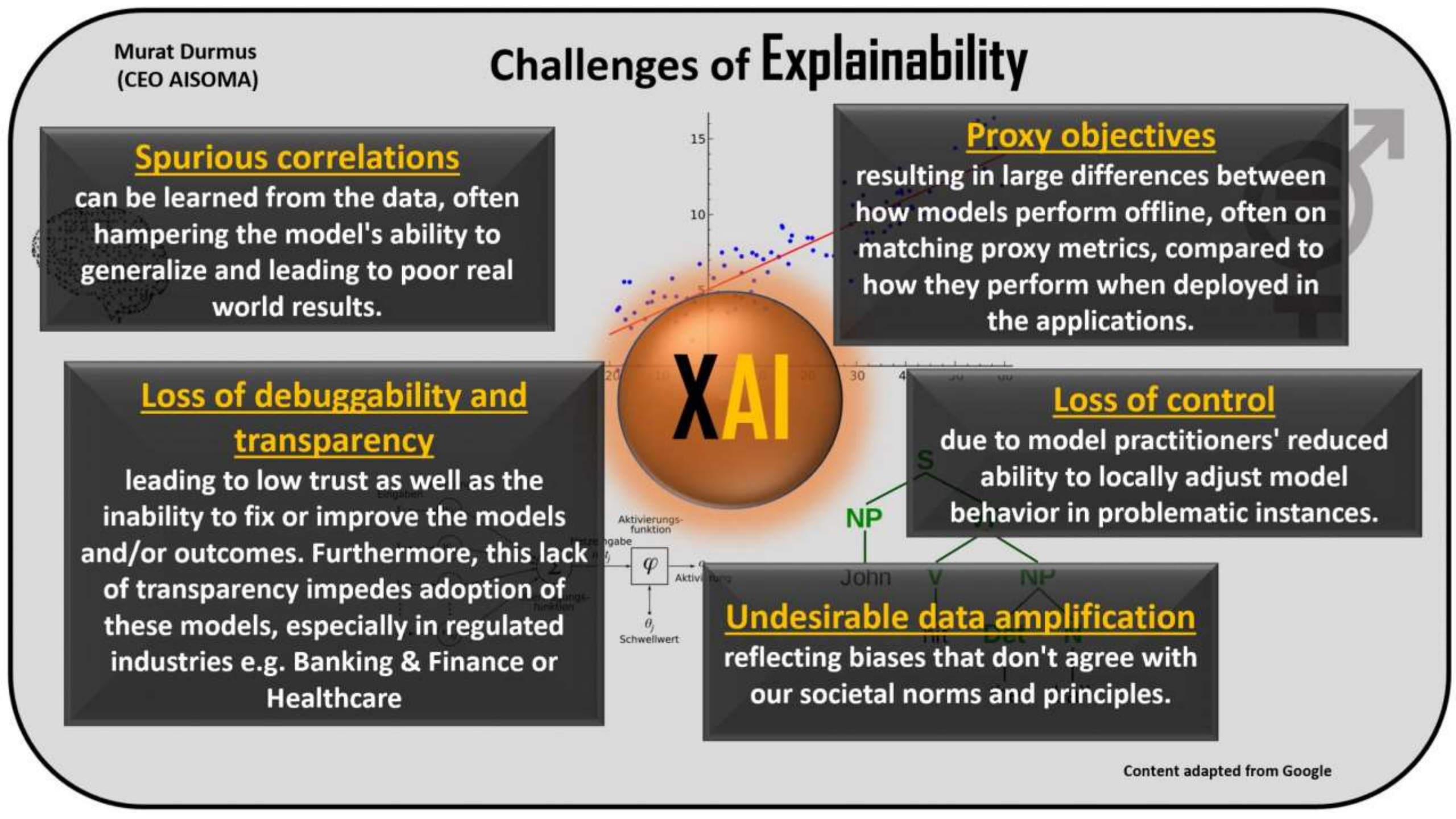
Challenges of Explainability

Spurious correlations

can be learned from the data, often hampering the model's ability to generalize and leading to poor real world results.

Proxy objectives

resulting in large differences between how models perform offline, often on matching proxy metrics, compared to how they perform when deployed in the applications.



Loss of debuggability and transparency

leading to low trust as well as the inability to fix or improve the models and/or outcomes. Furthermore, this lack of transparency impedes adoption of these models, especially in regulated industries e.g. Banking & Finance or Healthcare

Loss of control

due to model practitioners' reduced ability to locally adjust model behavior in problematic instances.

Undesirable data amplification

reflecting biases that don't agree with our societal norms and principles.

Content adapted from Google

Murat Durmus
(CEO AISOMA)

Main ethical and moral issues associated with the development and implementation of AI

Impact on society

- The labour market
- Inequality
- Privacy, human rights and dignity
- Bias
- Democracy

Impact on human psychology

- Relationships

Impact on the financial system

- Risk Management
- Fraud detection
- Automation

Impact on the legal system

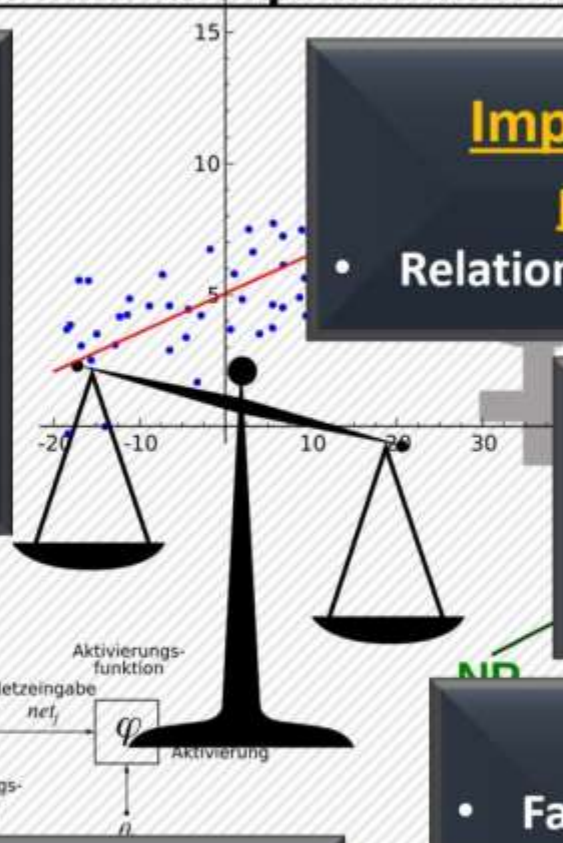
- Criminal law
- Tort law

Impact on the planet and environment

- Sustainability

Impact on trust

- Fairness
- Transparency
- Accountability
- Control



EPRS | European Parliamentary Research Service

Five ethical challenges of Artificial Intelligence

aisoma.de

rules

wrongs

morality

ideals

moral code

principles

rights

The Ethics of Artificial Intelligence

is the part of the ethics of technology specific to robots and other artificially intelligent beings. It is typically [citation needed] divided into roboethics, a concern with the moral behavior of humans as they design, construct, use and treat artificially intelligent beings, and machine ethics, which is concerned with the moral behavior of artificial moral agents (AMAs). *Wikipedia*

Created by Murat Durmus (CEO AISOMA)

Created by Murat Durmus (CEO AISOMA)

According to the European Ethics Guidelines for Trustworthy AI

Requirements of Trustworthy AI

1 Human agency and oversight

Including fundamental rights, human agency and human oversight.

2 Technical robustness and safety

Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.

3 Privacy and data governance

Including respect for privacy, quality and integrity of data, and access to data.

4 Transparency

Including traceability, explainability and communication.

5 Diversity, non-discrimination and fairness

Including the avoidance of unfair bias, accessibility & universal design, & stakeholder participation

6 Societal and environmental wellbeing

Including sustainability & environmental friendliness, social impact, society & democracy

7 Accountability

Including auditability, minimisation & reporting of negative impact, trade-offs & redress

To be continuously evaluated and addressed throughout the AI system's Life Cycle

Created by Murat Durmus (CEO AISOMA)

aisoma.de

Artificial Intelligence: 4 Ethical Implications

(According to The World Economic Forum)

1. Active Inclusion: the development and design of machine learning applications must actively seek a diversity of input, especially of the norms and values of specific populations affected by the output of AI systems

2. Fairness: People involved in conceptualizing, developing, and implementing machine learning systems should consider which definition of fairness best applies to their context and application, and prioritize it in the architecture of the machine learning system and its evaluation metrics.

3. Right to Understanding: Involvement of machine learning systems in decision-making that affects individual rights must be disclosed, and the systems must be able to provide an explanation of their decision-making that is understandable to end users and reviewable by a competent human authority. Where this is impossible and rights are at stake, leaders in the design, deployment, and regulation of machine learning technology must question whether or not it should be used

4. Access to Redress: Leaders, designers, and developers of machine learning systems are responsible for identifying the potential negative human rights impacts of their systems. They must make visible avenues for redress for those affected by disparate impacts, and establish processes for the timely redress of any discriminatory outputs.

An overview of model agnostic, local, and post hoc Interpretability Methods

Murat Durmus
(CEO AISOMA)

Method	Description	Pros	Cons
LIME	LIME, or Local Interpretable Model-Agnostic Explanations, is an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.	- Fully model-agnostic - Many contributors	- The inclusion of unrealistic data instances - Ambiguity of how to select kernel width - Coverage not always clear
SHAP	SHAP, or SHapley Additive exPlanations, is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.	-It can be seen how both functions move predictions up & down - Based on solid theory. SHAP values have useful and proven mathematical properties - The TreeExplainer is fast	- The inclusion of unrealistic data instances - Computationally expensive in the general case - Based on solid theory. It can be hard to dive into all the math inside this method
Permutation Importance	Permutation Importance is an intuitive way to assess the impact of a feature on the black-box model performance.	- Simple and intuitive - Available through the eli5 library (Python) - Easy to compute	- Labeled test data to compute the loss are required - Different shuffles may give different results - Greatly influenced by correlated features
Partial Dependency Plot	One of the simplest and most understandable methods is a partial dependency graph. This dependency graph shows us the dependency of a target variable from a particular feature.	- Easy and intuitive - Available in sklearn (Python)	- Assumption of feature independence - Loss of higher order interactions
Anchor	A novel model-agnostic system that explained the behavior of complex models with high-precision rules called anchors, representing local, “sufficient” conditions for predictions.	- Clear coverage - High precision - The anchors are expressive	- The inclusion of unrealistic data instances - The code available in the original repository is still in progress

Murat Durmus
(CEO AISOMA)

5 Themes to Consider for Ethical Data Science

Seek to enhance the value of data science for society	As the impact that data science can have on society could be significant, an important ethical consideration is what the potential implications could be on society as a whole.
Avoid harm	Data science has the potential to cause harm and this ethical consideration therefore focuses on how practitioners can avoid this by working in a manner that respects the privacy, equality and autonomy of individuals and groups, and speaking up about potential harm or ethical violations.
Apply and maintain professional competence	This ethical principle expects data science practitioners to apply best practice and comply with all relevant legal and regulatory requirements, as well as applicable professional body codes.
Seek to preserve or increase trustworthiness	The public's trust and confidence in the work of data scientists can be affected by the way ethical principles are applied. Practitioners can help to increase the trustworthiness of their work by considering ethical principles throughout all stages of a project.
Maintain accountability and oversight	Another key issue in data ethics around automation and AI is the question of how practitioners maintain human accountability and oversight within their work.

adopted by the Royal Statistical Society (RSS) and the Institute and Faculty of Actuaries (IFoA)

Psychological Model of Explanation (IHMC)

Murat Durmus
(CEO AISOMA)

XAI System

Model of the Explanation

Process and Possible Metrics

User

receives

Explanation

revises

User's Mental Model

enables

Better Performance

may initially

Is assessed by

Is assessed by

Is assessed by

Goodness Criteria

Test of Satisfaction

Test of Comprehension

Test of Performance

involves

can engender

Trust or Mistrust

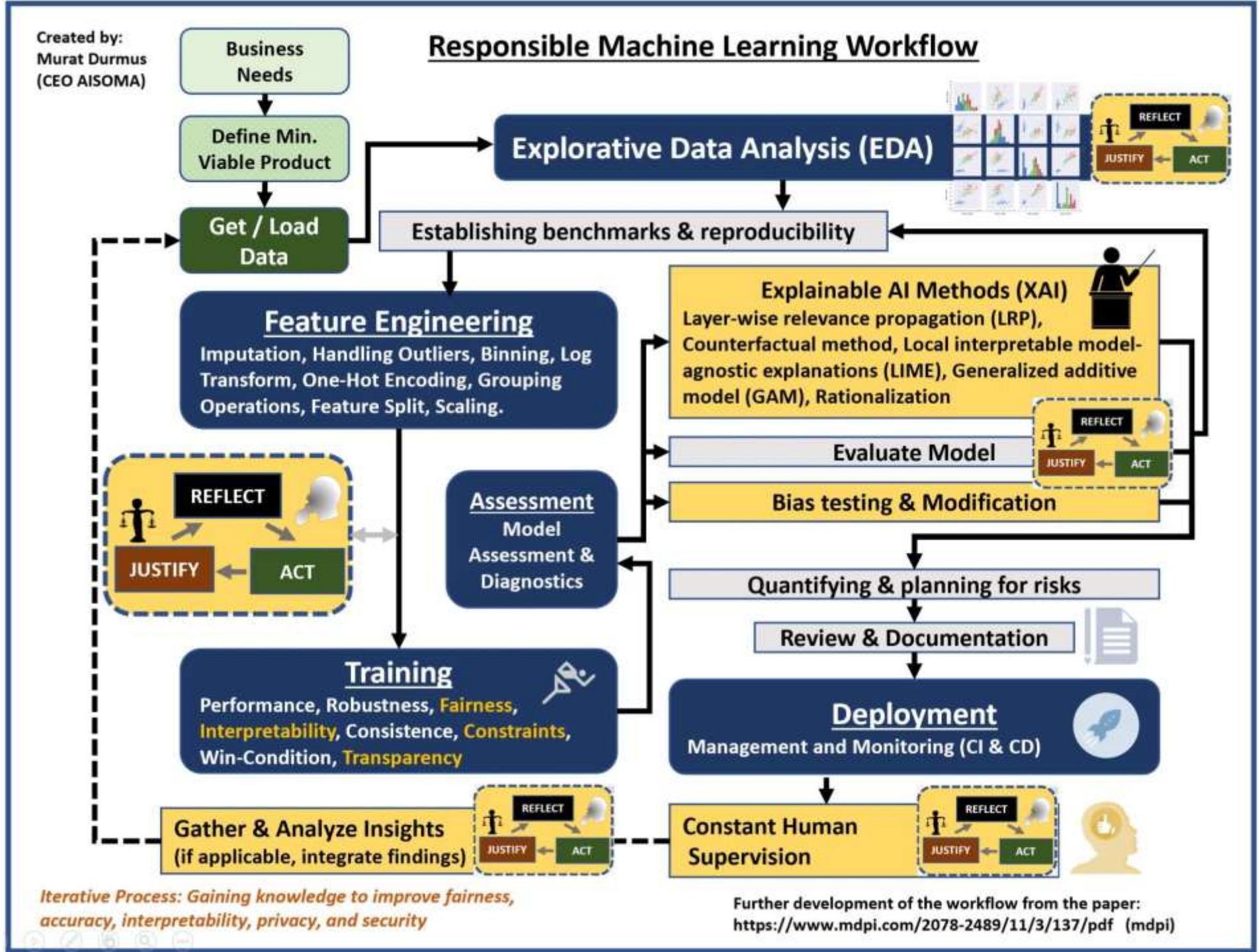
gives away to

Appropriate Trust

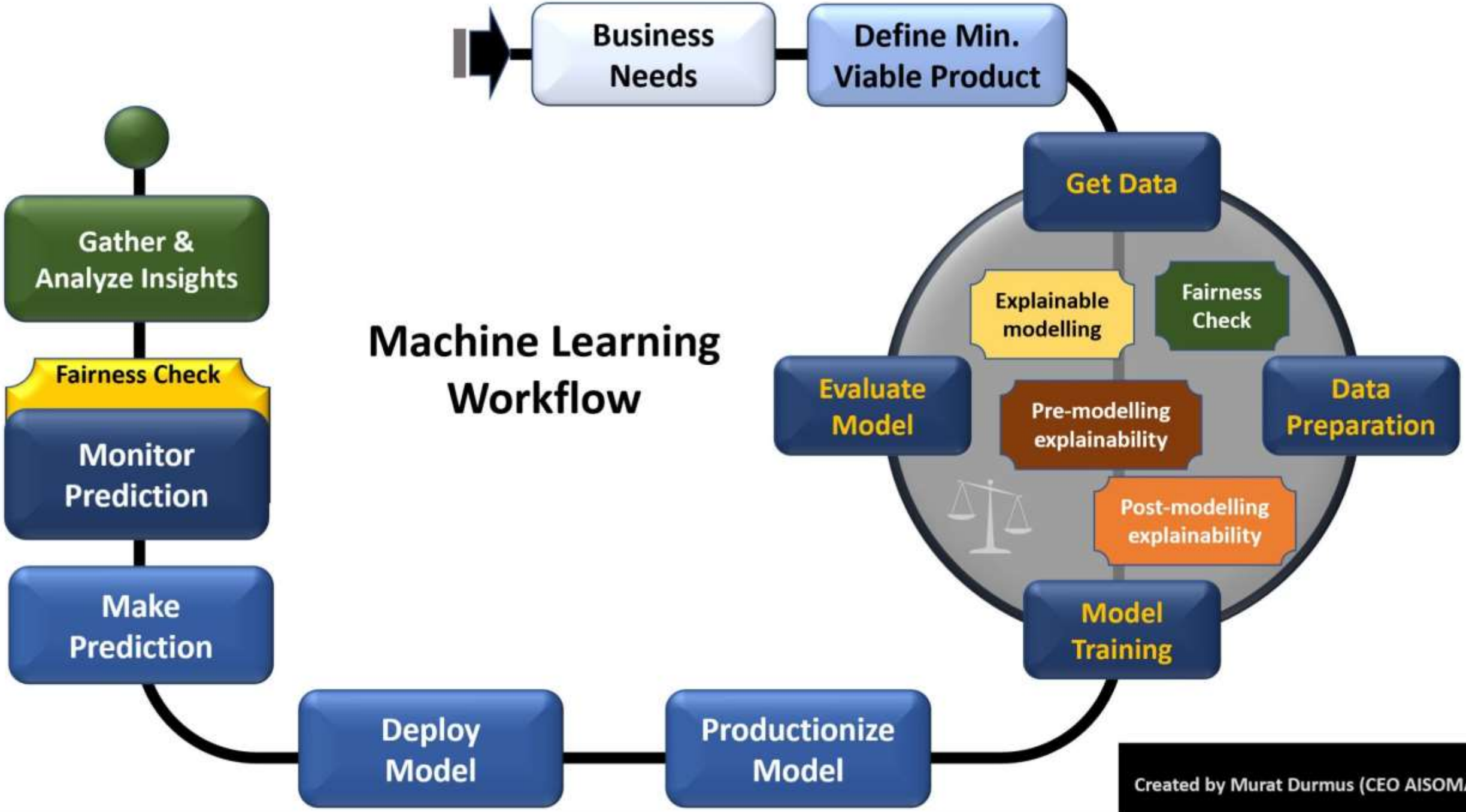
enables

Appropriate Use

Content Source: DARPA


Created by:
Murat Durmus
(CEO AISOMA)
Business Needs
Define Min. Viable Product
Get / Load Data
Responsible Machine Learning Workflow
Explorative Data Analysis (EDA)
REFLECT
JUSTIFY
ACT
Establishing benchmarks & reproducibility
Feature Engineering
Imputation, Handling Outliers, Binning, Log Transform, One-Hot Encoding, Grouping Operations, Feature Split, Scaling.
Explainable AI Methods (XAI)
Layer-wise relevance propagation (LRP), Counterfactual method, Local interpretable model-agnostic explanations (LIME), Generalized additive model (GAM), Rationalization
Evaluate Model
Bias testing & Modification
Assessment
Model Assessment & Diagnostics
Quantifying & planning for risks
Review & Documentation
Training
Performance, Robustness, Fairness, Interpretability, Consistence, Constraints, Win-Condition, Transparency
Deployment
Management and Monitoring (CI & CD)
Gather & Analyze Insights
(if applicable, integrate findings)
Constant Human Supervision
Iterative Process: Gaining knowledge to improve fairness, accuracy, interpretability, privacy, and security
Further development of the workflow from the paper:
https://www.mdpi.com/2078-2489/11/3/137/pdf (mdpi)

Machine Learning Workflow

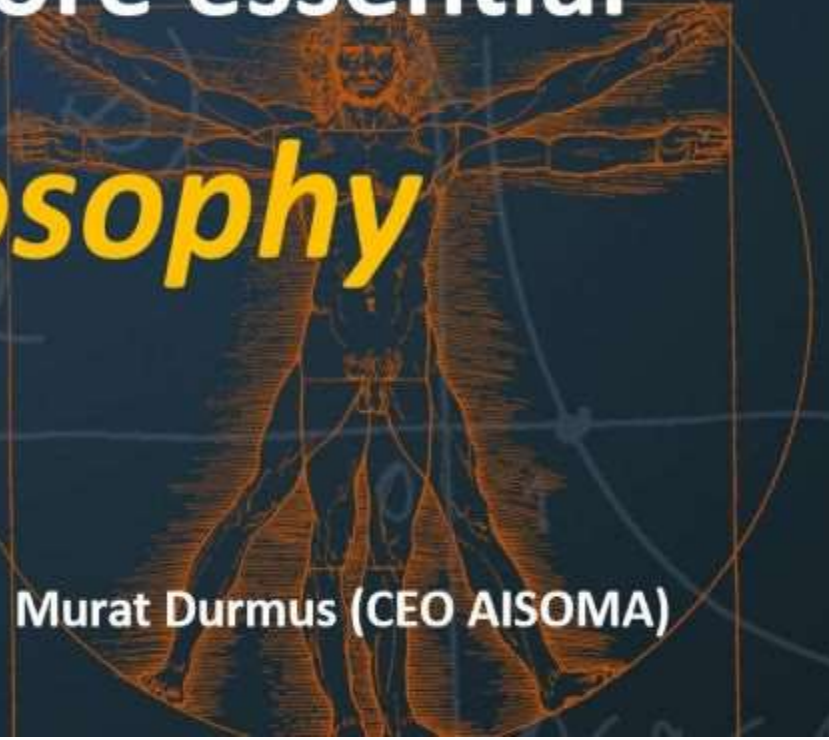


Created by Murat Durmus (CEO AISOMA)

Many are concerned about
the lack of *AI-Experts*.
The lack of "real" *Thinkers* & *Philosophers*
is even more alarming.

Murat Durmus
(CEO AISOMA)

The more *Artificial Intelligence* enters our lives, the more essential *Ethics & Philosophy* become.



Murat Durmus (CEO AISOMA)

aisoma.de

Contact

Murat Durmus

CEO & Founder @ AISOMA AG

Frankfurt am Main, Hessen, Deutschland

https://www.linkedin.com/in/ceosaisoma/

murat.durmus@aisoma.de

https://www.aisoma.de

Created by Murat Durmus (CEO AISOMA)