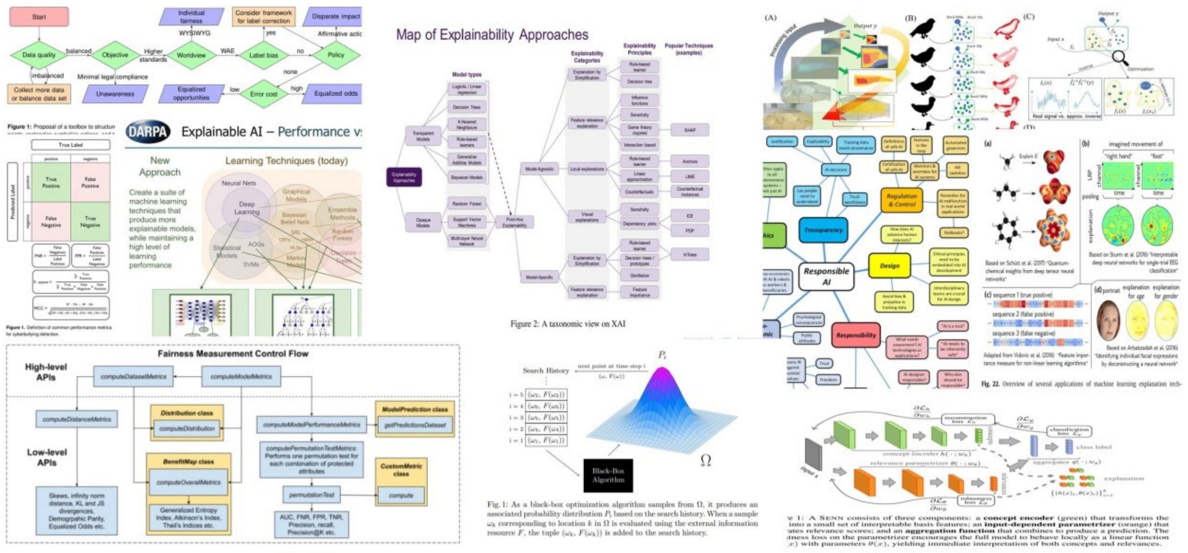
Created by Murat Durmus (CEO AISOMA)
LinkedIn: https://www.linkedin.com/in/ceosaisoma/

# A collection of recommendable papers and articles on Explainable AI (XAI)

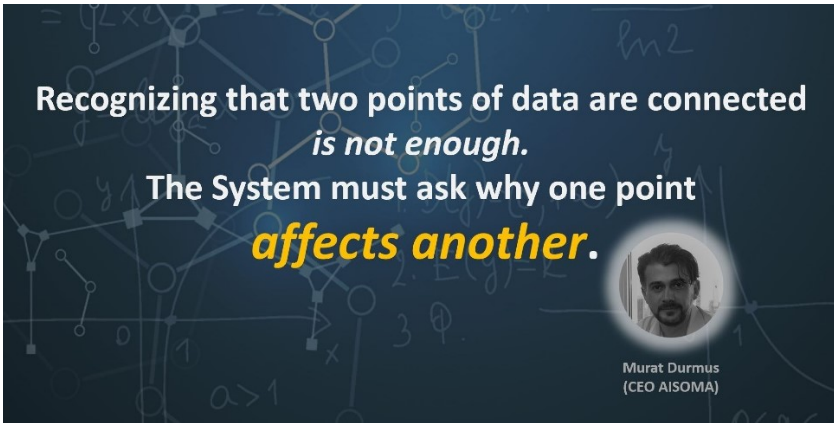

Map of Explainability Approaches
DARPA
Explainable AI – Performance vs
New Approach
Learning Techniques (today)
Fairness Measurement Control Flow
High-level APIs
Low-level APIs
Figure 2: A taxonomic view on XAI
Responsible AI
Transparency
Regulation & Control
Design
Responsibility
Black-Box Algorithm
Search History

## Content

## Intoduction

Explainable AI (XAI) refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by humans. It contrasts with the concept of the "black box" in machine learning where even their designers cannot explain why the AI arrived at a specific decision. XAI may be an implementation of the social right to explanation. XAI is relevant even if there is no legal rights or regulatory requirements—for example, XAI can improve the user experience of a product or service by helping end users trust that the AI is making good decisions.

The technical challenge of explaining AI decisions is sometimes known as the interpretability problem. Another consideration is infobesity (overload of information), thus, full transparency may not be always possible or even required. However, simplification at the cost of misleading users in order to increase trust or to hide undesirable attributes of the system should be avoided by allowing a tradeoff between interpretability and completeness of an explanation. (more info: ***Wikipedia***)


Recognizing that two points of data are connected
*is not enough.*
The System must ask why one point
*affects another.*
Murat Durmus
(CEO AISOMA)

Following a collection of recommendable papers and articles on Explainable AI (XAI):

### Great overview/introduction: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.

In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in front of the barrier of explainability, an inherent problem of the latest techniques brought by sub-symbolism (e.g. ensembles or Deep Neural Networks) that were not present in the last hype of AI (namely, expert systems and rule based models). Paradigms underlying this problem fall within the so-called eXplainable AI (XAI) field, which is widely acknowledged as a crucial feature for the practical deployment of AI models. The overview presented in this article examines the existing literature and contributions already done in the field of XAI, including a prospect toward what is yet to be reached.

Source(pdf): **https://lnkd.in/dDNDKz9**

### Definitions, methods, and applications in interpretable machine learning

Machine-learning models have demonstrated great success in learning complex patterns that enable them to make predictions about unobserved data. In addition to using models for prediction, the ability to interpret what a model has learned is receiving an increasing amount of attention. However, this increased focus has led to considerable confusion about the notion of interpretability. In particular, it is unclear how the wide array of proposed interpretation methods are related and what common concepts can be used to evaluate them

Source(pdf): **https://lnkd.in/dpBhynJ**

### Explainable and privacy-preserving artificial intelligence

Machine learning (ML) affects data privacy in two ways. It may be using sensitive personal data for training the models and (as ML models accuracy generally rises with amount of training data, the more data the better) and secondly, it may be affecting data privacy is when they are part of making decisions about humans.

Source(pdf): **https://lnkd.in/d29yc3r**

## explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning

Since the first presentation of neural networks in the 1940s , we have seen a great increase in works on Artificial Intelligence (AI) and Machine Learning (ML). Especially within the last decade, computational resources have become cheaper and more accessible. This development has led to new state-of-the-art solutions, e.g., Deep Learning (DL), while the increasing availability of tools and libraries has led to a democratization of ML methods in a variety of domains [30]. For example, DL methods outperform traditional algorithms for image processing [56] or natural language processing and can often be applied by domain experts without prior ML expertise.

Source(pdf): **https://lnkd.in/dTRK_MA**

## Getting Fairness Right: Towards a Toolbox for Practitioners

The potential risk of AI systems unintentionally embedding and reproducing bias has attracted the attention of machine learning practitioners and society at large. As policy makers are willing to set the standards of algorithms and AI techniques, the issue on how to refine existing regulation, in order to enforce that decisions made by automated systems are fair and non-discriminatory, is again critical. Meanwhile, researchers have demonstrated that the various existing metrics for fairness are statistically mutually exclusive and the right choice mostly depends on the use case and the definition of fairness.

Source(pdf): **https://lnkd.in/duugmjz**

## Formalizing Trust in Artificial Intelligence - Prerequisites, Causes and Goals of Human Trust in AI

Trust is a central component of the interaction between people and AI, in that ’incorrect’ levels of trust may cause misuse, abuse or disuse of the technology. But what, precisely, is the nature of trust in AI? What are the prerequisites and goals of the cognitive mechanism of trust, and how can we cause these prerequisites and goals, or assess whether they are being satisfied in a given interaction? This work aims to answer these questions. We discuss a model of trust inspired by, but not identical to, sociology’s interpersonal trust (i.e., trust between people). This model rests on two key properties of the vulnerability of the user and the ability to anticipate the impact of the AI model’s decisions. We incorporate a formalization of ‘contractual trust’, such that trust between a user and an AI is trust that some implicit or explicit contract will hold, and a formalization of ‘trustworthiness’ (which detaches from the notion of trustworthiness in sociology), and with it concepts of ’warranted’ and ‘unwarranted’ trust. We then present the possible causes of warranted trust as intrinsic reasoning and extrinsic behavior, and discuss how to design trustworthy AI, how to evaluate whether trust has manifested, and whether it is warranted. Finally, we elucidate the connection between trust and XAI using our formalization.

Source(pdf): **https://lnkd.in/dk-JyZb**

### A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI

Recently, artificial intelligence and machine learning in general have demonstrated remarkable performances in many tasks, from image processing to natural language processing, especially with the advent of deep learning. Along with research progress, they have encroached upon many different fields and disciplines. Some of them require high level of accountability and thus transparency, for example the medical sector. Explanations for machine decisions and predictions are thus needed to justify their reliability. This requires greater interpretability, which often means we need to understand the mechanism underlying the algorithms.

Source(pdf): **https://lnkd.in/dAu_Kqq**

### A Framework for Understanding Unintended Consequences of Machine Learning

As machine learning increasingly affects people and society, it is important that we strive for a comprehensive and unified understanding of potential sources of unwanted consequences. For instance, downstream harms to particular groups are often blamed on “biased data,” but this concept encompass too many issues to be useful in developing solutions. In this paper, we provide a framework that partitions sources of downstream harm in machine learning into six distinct categories spanning the data generation and machine learning pipeline. We describe how these issues arise, how they are relevant to particular applications, and how they motivate different solutions. In doing so, we aim to facilitate the development of solutions that stem from an understanding of application-specific populations and data generation processes, rather than relying on general statements about what may or may not be “fair.”

Source(pdf): **https://lnkd.in/dRqq_qa**

### Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices

Large and ever-evolving technology companies continue to invest more time and resources to incorporate responsible Artificial Intelligence (AI) into production-ready systems to increase algorithmic accountability. This paper examines and seeks to offer a framework for analyzing how organizational culture and structure impact the effectiveness of responsible AI initiatives in practice. We present the results of semi-structured qualitative interviews with practitioners working in industry, investigating common challenges, ethical tensions, and effective enablers for responsible AI initiatives. Focusing on major companies developing or utilizing AI, we have mapped what organizational structures currently support or hinder responsible AI initiatives, what aspirational future processes and structures would best enable effective initiatives, and what key elements comprise the transition from current work practices to the aspirational future.

Source(pdf): **https://lnkd.in/dY7trBh**

## The Frutility of Bias-Free Learning and Search

Building on the view of machine learning as search, we demonstrate the necessity of bias in learning, quantifying the role of bias (measured relative to a collection of possible datasets, or more generally, information resources) in increasing the probability of success. For a given degree of bias towards a fixed target, we show that the proportion of favorable information resources is strictly bounded from above. Furthermore, we demonstrate that bias is a conserved quantity, such that no algorithm can be favorably biased towards many distinct targets simultaneously. Thus bias encodes trade-offs. The probability of success for a task can also be measured geometrically, as the angle of agreement between what holds for the actual task and what is assumed by the algorithm, represented in its bias. Lastly, finding a favorably biasing distribution over a fixed set of information resources is provably difficult, unless the set of resources itself is already favorable with respect to the given task and algorithm.

Source(pdf): **https://lnkd.in/dK4NGRD**

## Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition

In 2018, a landmark challenge in artificial intelligence (AI) took place, namely, the Explainable Machine Learning Challenge. The goal of the competition was to create a complicated black box model for the dataset and explain how it worked. One team did not follow the rules. Instead of sending in a black box, they created a model that was fully interpretable. This leads to the question of whether the real world of machine learning is similar to the Explainable Machine Learning Challenge, where black box models are used even when they are not needed. We discuss this team's thought processes during the competition and their implications, which reach far beyond the competition itself.

Source: **https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/5**

## Responsible AI – Key Themes, Concerns & Recommendations for European Research and Innovation

This document's purpose is to provide input into the advisory processes that determine European support for both research into Responsible AI; and how innovation using AI that takes into account issues of responsibility can be supported. "Responsible AI" is an umbrella term for investigations into legal, ethical and moral standpoints of autonomous algorithms or applications of AI whose actions may be safetycritical or impact the lives of citizens in significant and disruptive ways.

Source(pdf): **https://lnkd.in/gD4FH5k**

## Self-explaining AI as an alternative to interpretable AI

While it is often possible to approximate the inputoutput relations of deep neural networks with a few human-understandable rules, the discovery of the double descent phenomena suggests that such approximations do not accurately capture the mechanism by which deep neural networks work. Double descent indicates that deep neural networks typically operate by smoothly interpolating between data points rather than by extracting a few high level rules. As a result, neural networks trained on complex real world data are inherently hard to interpret and prone to failure if asked to extrapolate. To show how we might be able to trust AI despite these problems we explore the concept of self-explaining AI, which provides both a prediction and explanation. We also argue AIs systems should include a "warning light" using techniques from applicability domain analysis and anomaly detection to warn the user if a model is asked to extrapolate outside its training distribution.

Source(pdf): **https://lnkd.in/d5hEn6d**

## Principles and Practice of Explainable Machine Learning

Artificial intelligence (AI) provides many opportunities to improve private and public life. Discovering patterns and structures in large troves of data in an automated manner is a core component of data science, and currently drives applications in diverse areas such as computational biology, law and finance. However, such a highly positive impact is coupled with significant challenges: how do we understand the decisions suggested by these systems in order that we can trust them? In this report, we focus specifically on data-driven methods – machine learning (ML) and pattern recognition models in particular – so as to survey and distill the results and observations from the literature.

Source(pdf): **https://lnkd.in/dzgKaBn**

### A Survey of the State of Explainable AI for Natural Language Processing

Recent years have seen important advances in the quality of state-of-the-art models, but this has come at the expense of models becoming less interpretable. This survey presents an overview of the current state of Explainable AI (XAI), considered within the domain of Natural Language Processing (NLP). We discuss the main categorization of explanations, as well as the various ways explanations can be arrived at and visualized. We detail the operations and explainability techniques currently available for generating explanations for NLP model predictions, to serve as a resource for model developers in the community. Finally, we point out the current gaps and encourage directions for future work in this important research area

Source(pdf): **https://lnkd.in/d3CPt7s**

### Towards Robust Interpretability with Self-Explaining Neural Networks

Most recent work on interpretability of complex machine learning models has focused on estimating a posteriori explanations for previously trained models around specific predictions. Self-explaining models where interpretability plays a key role already during learning have received much less attention. We propose three desiderata for explanations in general – explicitness, faithfulness, and stability – and show that existing methods do not satisfy them. In response, we design self-explaining models in stages, progressively generalizing linear classifiers to complex yet architecturally explicit models. Faithfulness and stability are enforced via regularization specifically tailored to such models. Experimental results across various benchmark datasets show that our framework offers a promising direction for reconciling model complexity and interpretability.

Source(pdf): **https://lnkd.in/ds9JdaG**

- **An overview of some available Fairness Frameworks & Packages**

  LinkedIn Pulse article: **https://www.linkedin.com/pulse/overview-some-available-fairness-frameworks-packages-murat-durmus/**

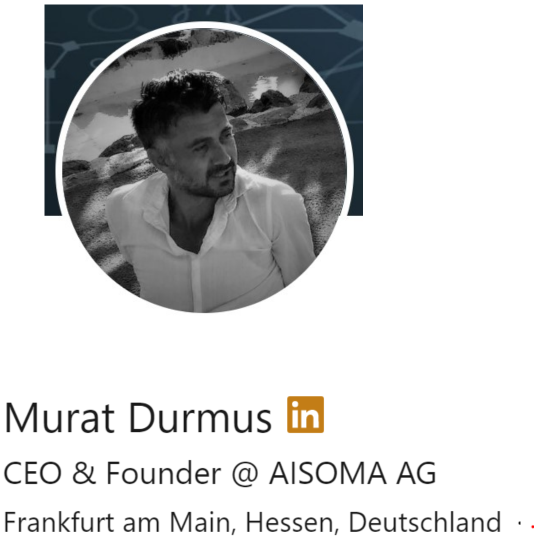
- **Inside the Black Box: 5 Methods for Explainable-AI**

  LinkedIn Pulse article: **https://bit.ly/322BHIV**

- **Slides & quotes on AI-Ethics & XAI**:

  LinkedIn Pulse article: **https://lnkd.in/dYxGrec**

# 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