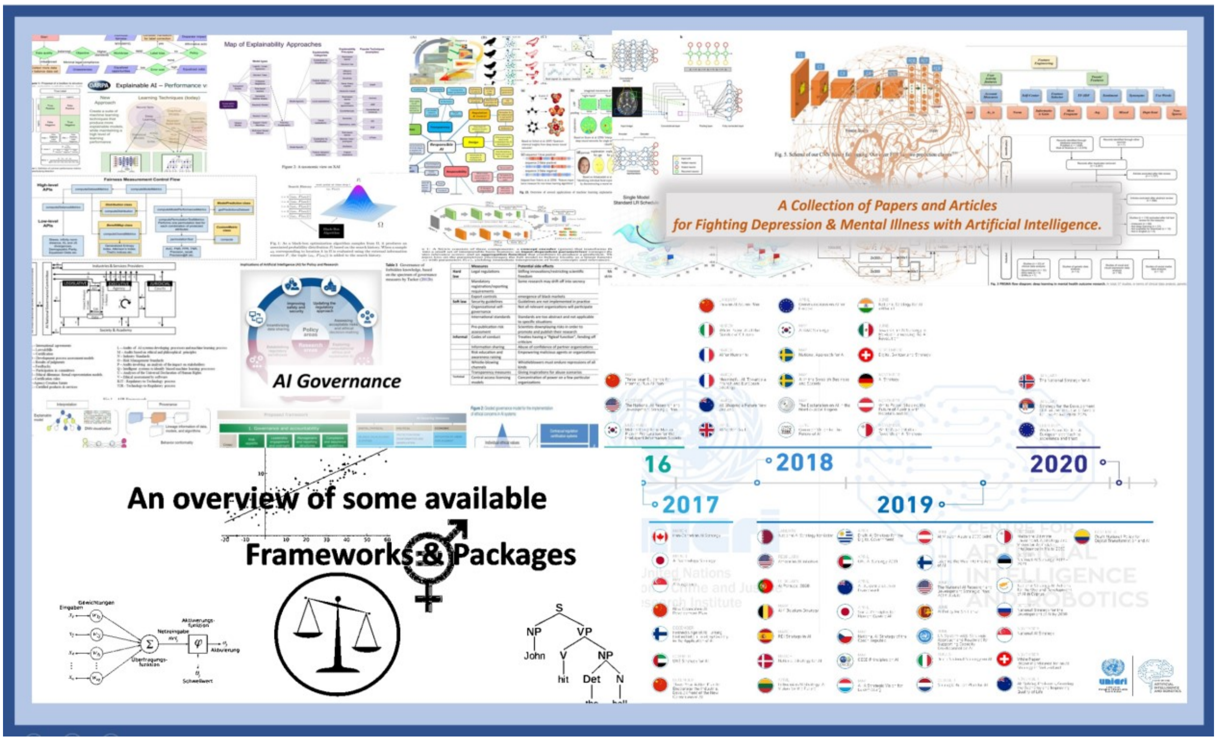
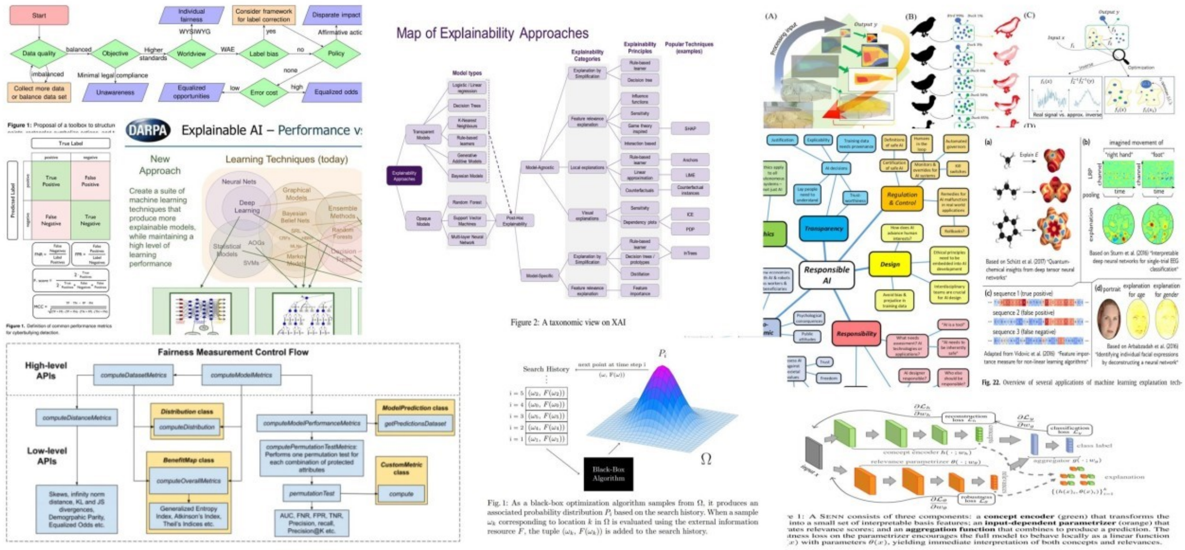
Created by Murat Durmus (CEO AISOMA)
LinkedIn: https://www.linkedin.com/in/ceosaisoma/

The Full Compilation 2020

A Collection of Recommendable Papers & Articles on AI, XAI, AI-Ethics, AI-Governance, Regulation, AI for Mental-Health, and Well-Being


Map of Explainability Approaches
Explainable AI – Performance vs.
A Collection of Papers and Articles
for Fighting Depression & Mental Illness with Artificial Intelligence.
AI Governance
An overview of some available
Frameworks & Packages
16
2017
2018
2019
2020
NP
VP
John
V
hit
Det
N
S

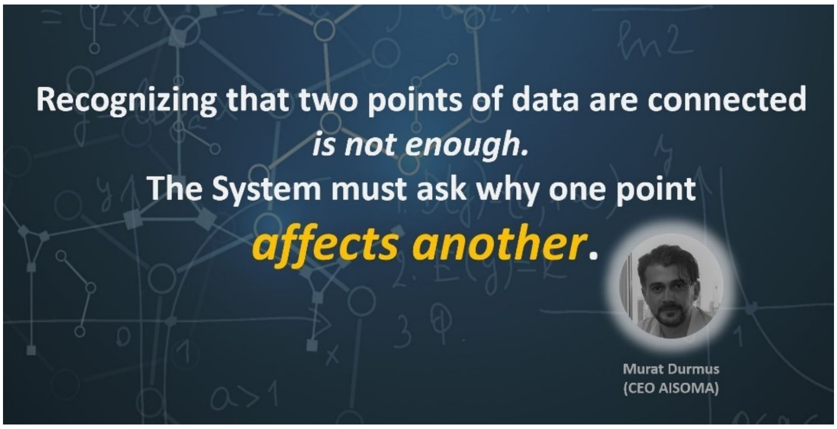
A collection of recommendable papers and articles on Explainable AI (XAI)


DARPA Explainable AI – Performance vs
Map of Explainability Approaches
Figure 2: A taxonomic view on XAI
Fairness Measurement Control Flow
High-level APIs
Low-level APIs
Black-Box Algorithm

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**

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

A Collection of Papers and Articles for *Fighting Depression* & Mental Illness with Artificial Intelligence.

A Collection of Papers and Articles
for Fighting Depression & Mental Illness with Artificial Intelligence.
Fig. 7. Cyclic stochastic gradient descent with restart (SGDR): (a) cycle_len = 1; (b) cycle_len = 3 [26].
Fig. 3. Network architecture used for the convolutional neural network models described in this paper. Adapted from [30].
Fig. 5. Schema of our CNN during fine-tuning. Our layer FC8 has two prediction classes
Fig. 3 PRISMA flow diagram: deep learning in mental health outcome research.
Single Model
Standard LR Schedule
Feature Engineering
User Activity features
Tweets' Features
Account Measures
Self-Center
Feature Selector
TF-IDF
Sentiment
Synonyms
Use Words
freeze layers
train

Inhaltsverzeichnis

Mental illness is a grave issue. We must not underestimate and downplay it because the future of humanity depends on it. ~ Murat

Depressed, mentally unstable people suffer, especially in these times of social distancing. Even for healthy people, this time is a big challenge. We should increasingly develop technological possibilities and put them at the top of our list of problems to be solved. There is much talk of AI4Good. This is an area where AI can alleviate and even cure people's ever-increasing mental illness. We need to accelerate the transfer from research to practice and prioritize and promote significant mental illness, climate, sustainable and ethical business, inequality, etc.

We look into an uncertain future with many challenges. We can only shape this future positively if we face it with a healthy and stable spirit. ~ Murat

*"Mental health, defined by the World Health Organization (WHO), is "a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community". The three core components of this definition are (1) well-being, (2) effective functioning of an individual, and effective functioning for a community. According to the WHO , mental health includes "subjective well-being, perceived self-efficacy, autonomy, competence, intergenerational dependence, and self-actualization of one's intellectual and emotional potential, among others". From the perspectives of positive psychology or of holism, mental health may include an individual's ability to enjoy life and to create a balance between life activities and efforts to achieve psychological resilience. Cultural differences, subjective assessments, and competing professional theories all affect how one defines "mental health" (source: **wikipedia**)*

According to **World Economic Forum**:

"Depression and anxiety disorders cost the global economy $1 trillion every year in lost productivity - and take a terrible human toll. " ~ WEF

In the following a collection of Papers/Articles for Fighting Depression & Mental Illness with Artificial Intelligence.

Papers

Machine Learning-based Approach for Depression Detection in Twitter Using Content and Activity Features

Social media channels, such as Facebook, Twitter, and Instagram, have altered our world forever. People are now increasingly connected than ever and reveal a sort of digital persona. Although social media certainly has several remarkable features, the demerits are undeniable as well. Recent studies have indicated a correlation between high usage of social media sites and increased depression. The present study aims to exploit machine learning techniques for detecting a probable depressed Twitter user based on both, his/her network behavior and tweets. For this purpose, we trained and tested classifiers to distinguish whether a user is depressed or not using features extracted from his/her activities in the network and tweets. The results showed that the more features are used, the higher are the accuracy and F-measure scores in detecting depressed users. This method is a data-driven, predictive approach for early detection of depression or other mental illnesses. This study's main contribution is the exploration part of the features and its impact on detecting the depression level

Source(pdf): **https://arxiv.org/ftp/arxiv/papers/2003/2003.04763.pdf**

Artificial Intelligence for Mental Health and Mental Illnesses: an Overview

Purpose of Review Artificial intelligence (AI) technology holds both great promise to transform mental healthcare and potential pitfalls. This article provides an overview of AI and current applications in healthcare, a review of recent original research on AI specific to mental health, and a discussion of how AI can supplement clinical practice while considering its current limitations, areas needing additional research, and ethical implications regarding AI technology

Source(pdf):https://escholarship.org/content/qt9gx593b0/qt9gx593b0_noSplash_d814b6b41c76cb874050695d2bf30ced.pdf

Methods in predictive techniques for mental health status on social media: a critical review

Researchers in computer science (CS) are using behavioral and linguistic cues from social media data to predict the presence of mood and psychosocial disorders. Since 2013, research can assess the presence of major depression1–3 , suicidality4–6 , eating disorders7,8 , and schizophrenia9 , among others with high accuracy (80–90%). In addition to mental disorders, these approaches are starting to assess related symptomatology, such as self-harm8 , stress10, and the severity of mental illness11 without the use of inperson, clinical assessment. These signals are taken from the posting and behavioral history of social media websites and apps, such as Twitter, Reddit, and Facebook12. In this article, we adopt the term mental health status (MHS) to capture both mental disorders and these related symptomatology

Source(pdf): **https://www.nature.com/articles/s41746-020-0233-7.pdf**

Deep learning in mental health outcome research: a scoping review

Mental illnesses, such as depression, are highly prevalent and have been shown to impact an individual's physical health. Recently, artificial intelligence (AI) methods have been introduced to assist mental health providers, including psychiatrists and psychologists, for decision-making based on patients' historical data (e.g., medical records, behavioral data, social media usage, etc.). Deep learning (DL), as one of the most recent generation of AI technologies, has demonstrated superior performance in many real-world applications ranging from computer vision to healthcare.

Source: **https://www.nature.com/articles/s41398-020-0780-3**

A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media

Powered by machine learning techniques, social media provides an unobtrusive lens into individual behaviors, emotions, and psychological states. Recent research has successfully employed social media data to predict mental health states of individuals, ranging from the presence and severity of mental disorders like depression to the risk of suicide. These algorithmic inferences hold great potential in supporting early detection and treatment of mental disorders and in the design of interventions. At the same time, the outcomes of this research can pose great risks to individuals, such as issues of incorrect, opaque algorithmic predictions, involvement of bad or unaccountable actors, and potential biases from intentional or inadvertent misuse of insights.

Source(pdf): **http://steviechancellor.com/wp-content/uploads/2019/03/taxonomy-prediction-mh-fat2019.pdf**

Automated speech-based screening of depression using deep convolutional neural networks

Early detection and treatment of depression is essential in promoting remission, preventing relapse, and reducing the emotional burden of the disease. Current diagnoses are primarily subjective, inconsistent across professionals, and expensive for individuals who may be in urgent need of help. This paper proposes a novel approach to automated depression detection in speech using convolutional neural network (CNN) and multipart interactive training. The model was tested using 2568 voice samples obtained from 77 non-depressed and 30 depressed individuals. In experiment conducted, data were applied to residual CNNs in the form of spectrograms—images auto-generated from audio samples. The experimental results obtained using different ResNet architectures gave a promising baseline accuracy reaching 77%

Source(pdf): **https://arxiv.org/ftp/arxiv/papers/1912/1912.01115.pdf**

Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences

Depression is ranked as the largest contributor to global disability and is also a major reason for suicide. Still, many individuals suffering from forms of depression are not treated for various reasons. Previous studies have shown that depression also has an effect on language usage and that many depressed individuals use social media platforms or the internet in general to get information or discuss their problems. This paper addresses the early detection of depression using machine learning models based on messages on a social platform. In particular, a convolutional neural network based on different word embeddings is evaluated and compared to a classification based on user-level linguistic metadata.

Source(pdf): **https://arxiv.org/pdf/1804.07000.pdf**

Detecting Depression Using a Framework Combining Deep Multimodal Neural Networks with a Purpose-Built Automated Evaluation

Machine learning (ML) has been introduced into the medical field as a means to provide diagnostic tools capable of enhancing accuracy and precision while minimizing laborious tasks that require human intervention. There is mounting evidence that the technology fueled by ML has the potential to detect, and substantially improve treatment of complex mental disorders such as depression. We developed a framework capable of detecting depression with minimal human intervention: AiME (Artificial Intelligence Mental Evaluation). AiME consists of a short human-computer interactive evaluation that utilizes artificial intelligence, namely deep learning, and can predict whether the participant is depressed or not with satisfactory performance. Due to its ease of use, this technology can offer a viable tool for mental health professionals to identify symptoms of depression, thus enabling a faster preventative intervention. Furthermore, it may alleviate the challenge of observing and interpreting highly nuanced physiological and behavioral biomarkers of depression by providing a more objective evaluation.

source(pdf): **https://aime-static.textpert.ai/pub/detecting-depression-using-ml.pdf**

A Collection of Papers & Articles on AI for Human Well-Being

Numerous AI initiatives are underway in the health sector. Some of these are aimed at promoting mental health and well-being. In the following, I would like to present some promising approaches/articles from this area.

source (LinkedIn pulse article): **https://www.linkedin.com/pulse/collection-papers-articles-ai-human-well-being-murat-durmus/**

Articles

- **Time: How Artificial Intelligence Can Help Pick the Best Depression Treatments for You**

https://time.com/5786081/depression-medication-treatment-artificial-intelligence/

- **Forbes: The Incredible Ways Artificial Intelligence Is Now Used In Mental Health**

https://www.forbes.com/sites/bernardmarr/2019/05/03/the-incredible-ways-artificial-intelligence-is-now-used-in-mental-health/?sh=6606e570d02e

- **Medium: Machine Learning in the Treatment of Depression**

https://medium.com/swlh/machine-learning-in-the-treatment-of-depression-87dcd63f528d

- **Science Node: Detecting depression with AI**

https://sciencenode.org/feature/Detecting%20depression.php

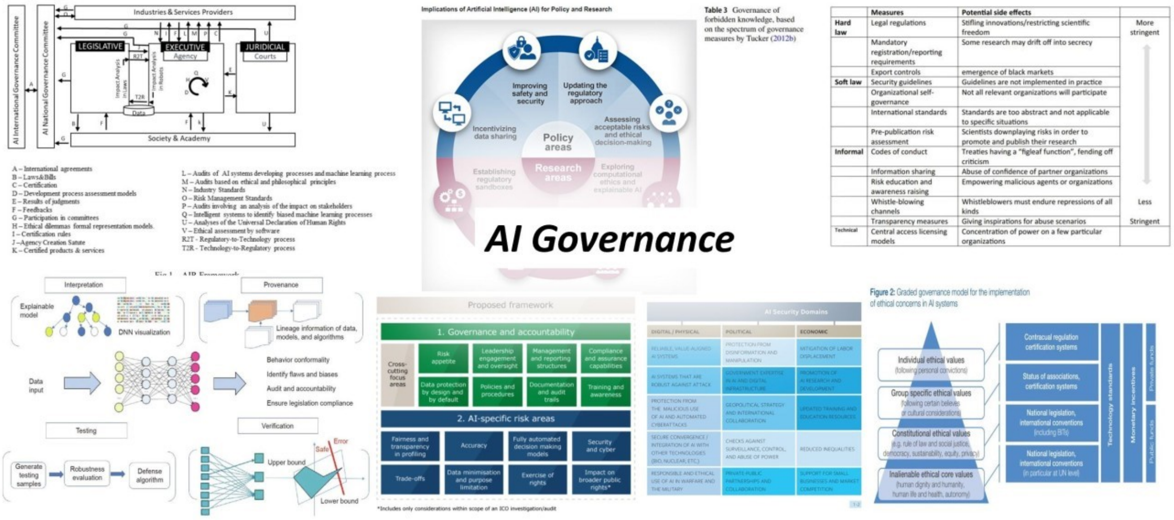
- **Health IT Analytics: Expanding Access to Mental Healthcare with Artificial Intelligence**

https://healthitanalytics.com/news/expanding-access-to-mental-healthcare-with-artificial-intelligence

- **Artificial Intelligence in Healthcare - Promising Progress (Best Use Cases)**

https://www.linkedin.com/pulse/artificial-intelligence-healthcare-promising-progress-murat-durmus/

A collection of recommendable papers and articles on *AI-Governance*


Industries & Services Providers
LEGISLATIVE
EXECUTIVE
JUDICIAL
Society & Academy
Implications of Artificial Intelligence (AI) for Policy and Research
Improving safety and security
Updating the regulatory approach
Incentivizing data sharing
Policy areas
Assessing acceptable risks and ethical decision-making
Establishing regulatory sandboxes
Research areas
Exploring computational ethics and explainable AI
AI Governance
Table 3 Governance of forbidden knowledge, based on the spectrum of governance measures by Tucker (2012b)
Hard law
Soft law
Informal
Technical
More stringent
Less Stringent
Interpretation
Provenance
Explainable model
DNN visualization
Lineage information of data, models, and algorithms
Data input
Behavior conformity
Identify flaws and biases
Audit and accountability
Ensure legislation compliance
Testing
Verification
Generate testing samples
Robustness evaluation
Defense algorithm
Upper bound
Lower bound
Proposed framework
1. Governance and accountability
2. AI-specific risk areas
AI Security Domains
Figure 2: Graded governance model for the implementation of ethical concerns in AI systems
Individual ethical values
Group specific ethical values
Constitutional ethical values
Inalienable ethical core values
Technology standards
Monetary incentives

The development and use of artificial intelligence (AI) technologies can bring about many benefits – from transforming businesses and improving labour productivity to enhancing quality of life. However, ***there is an increasing concern regarding the risk of harm associated with the use of AI technologies if they are not deployed in a responsible manner, and the data within these models is not managed properly***. Many governments and international organizations have worked to outline ethical principles to govern the development and use of new technologies, with the aim to mitigate the risk of harm that these technologies might bring. Singapore issued the first edition of its Model AI Governance Framework , a sector-, technology- and algorithm-agnostic framework, which converts relevant ethical principles to implementable practices in an AI deployment process so that organizations can operationalize these principles.(WEF)

Following a collection of recommendable papers and articles on AI-Governance

"Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance":

"This article presents a meta-framework for Artificial Intelligence (AI) regulation that encompasses all stages of international public policy-making, from formulation to sustainable governance. Based on a vast systematic review of the literature on Artificial Intelligence Regulation (AIR) published between 2009 and 2019, a dispersed body of knowledge organized under the label "framework" was identified, containing 15 unique frameworks and several different theories that created a complex scientific scenario for research and practice."

Source: **https://lnkd.in/dSz8Car**

It is important to be aware of this - a must-read - "Principles alone cannot guarantee ethical AI"

"Artificial intelligence (AI) ethics is now a global topic of discussion in academic and policy circles. At least 84 public–private initiatives have produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI. According to recent meta-analyses, AI ethics has seemingly converged on a set of principles that closely resemble the four classic principles of medical ethics. Despite the initial credibility granted to a principled approach to AI ethics by the connection to principles in medical ethics, there are reasons to be concerned about its future impact on AI development and governance."

source: **https://lnkd.in/dg3ZhDM**

Implications of AI for Policy and Research

As AI technologies continue to advance at an incredible speed, federal oversight considerations need to evolve alongside them.

Some questions for policymakers to consider when assessing these technologies include:

- How is the federal government using AI systems? For example, what data and code are used to power these technologies?

- How should AI systems be evaluated? What approaches should auditors take to develop credible assessments?

- What would an evidence-based AI assessment look like?

- What does the future hold for AI oversight?

source: **https://lnkd.in/dHWXVxr**

Great Paper on AI & Society: "Forbidden knowledge in machine learning refections on the limits of research and publication."

"Certain research strands can yield "forbidden knowledge". This term refers to knowledge that is considered too sensitive, dangerous or taboo to be produced or shared. Discourses about such publication restrictions are already entrenched in scientifc felds like IT security, synthetic biology or nuclear physics research. This paper makes the case for transferring this discourse to machine learning research. Some machine learning applications can very easily be misused and unfold harmful consequences, for instance, with regard to generative video or text synthesis, personality analysis, behavior manipulation, software vulnerability detection and the like"

source: **https://lnkd.in/dCWUcMw**

"An overview of the Auditing Framework for Artificial Intelligence and its core components"

Two key components will be discussed

1. governance and accountability; and

2. AI-specific risk areas.

The governance and accountability component will discuss the measures an organisation must have in place to be compliant with data protection requirements.

The second component will focus on the potential data protection risks that may arise in a number of AI specific areas and what the adequate risk management practices would be to manage them.

source: **https://lnkd.in/druJmuS**

Highly recommended: "Governing Artificial Intelligence: UPHOLDING HUMAN RIGHTS & DIGNITY"

"This report is intended as a resource for anyone working in the field of AI and governance. It is also intended for those in the human rights field, outlining why they should be concerned about the present-day impacts of AI. What follows translates between DATA & SOCIETY 2 GOVERNING ARTIFICIAL INTELLIGENCE these fields by reframing the societal impact of AI systems through the lens of human rights. As a starting point, we focus on five initial examples of human rights areas – nondiscrimination, equality, political participation, privacy, and freedom of expression – and demonstrate how each one is implicated in a number of recent controversies generated as a result of AI-related systems."

source: **https://lnkd.in/duWvpV5**

"Toward AI Security: Global Aspirations for a More Resilient Future"

"Artificial intelligence (AI) may be the most important global issue of the 21st century, and how we navigate the security implications of AI could dramatically shape the future.1 Although research in AI has been advancing since the 1950's, recent years have seen substantial growth in interest, investment dollars, and jobs in this field,2 leading to important advances in real-world applications ranging from autonomous vehicles to cancer screening.3 It has become clear that AI is a transformative general-purpose technology that will spread across geographies and sectors, resulting in massive potential benefits-and risks-that are difficult or impossible to foresee. This presents a set of coordination and cooperation challenges to firms, governments, and civil society organizations that are trying to understand and act, prospectively, to shape the evolution of AI for human benefit."

Executive Summary: **https://lnkd.in/dtte62C**

Full Report: **https://lnkd.in/diEwEDG**

Very important: "Practical approaches to implement ethics in AI systems"

"AI Governance: A Holistic Approach to Implement Ethics into AI" by World Economic Forum

There are many potential benefits to the application of Artificial Intelligence (AI) technologies, including the reduction of economic inefficiencies and increase in high-skilled jobs. There are also significant risks that must be managed — through both technical design and policy-making instruments— to maximize these benefits for any given society while protecting its important ethical values.

source: **https://lnkd.in/dhSkESk**

Highly recommended: A must-read for everyone interested in Governance and Ethics.

"Understanding artificial intelligence ethics and safety A guide for the responsible design and implementation of AI systems in the public sector"

This document provides end-to-end guidance on how to apply principles of AI ethics and safety to the design and implementation of algorithmic systems in the public sector. We will shortly release a workbook to bring the recommendations made in this guide to life. The workbook will contain case studies highlighting how the guidance contained here can be applied to concrete AI projects. It will also contain exercises and practical tools to help strengthen the process-based governance of your AI project.

source: **https://lnkd.in/dYzi3pN**

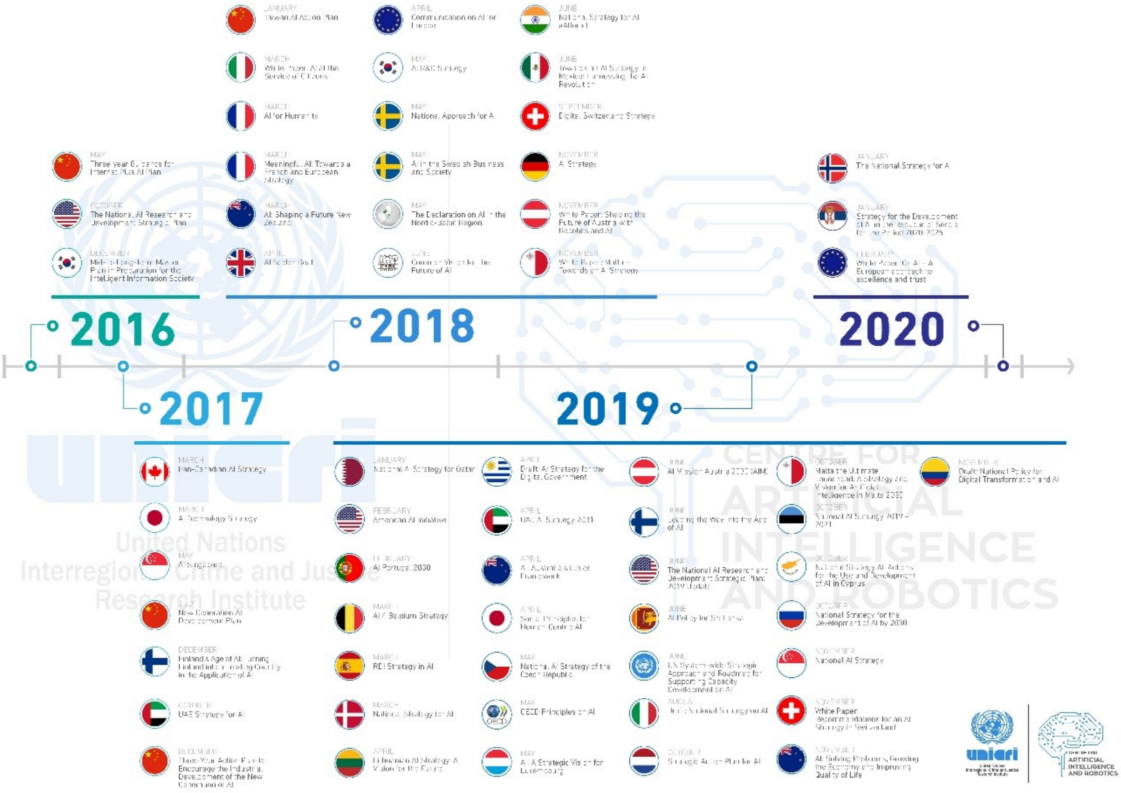
This might be also of interest:

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

An overview of some available Fairness Frameworks & Packages

A collection of useful Slides & Quotes on AI-Ethics and XAI

Regulation of Artificial Intelligence
Timeline & Strategies


2016
MAY
Three year Guidance for Internet Plus AI Plan
OCTOBER
The National AI Research and Development Strategic Plan
DECEMBER
Mid- to Long-term Master Plan in Preparation for the Intelligent Information Society
2017
MARCH
Pan-Canadian AI Strategy
MARCH
AI Technology Strategy
MAY
AI Singapore
JULY
New Generation AI Development Plan
DECEMBER
Finland's Age of AI: Turning Finland into a Leading Country in the Application of AI
OCTOBER
UAE Strategy for AI
DECEMBER
Three-Year Action Plan to Encourage the Industrial Development of the New Generation of AI
2018
JANUARY
Taiwan AI Action Plan
MARCH
White Paper: AI at the Service of Citizens
MARCH
AI for Humanity
MARCH
Meaningful AI: Towards a French and European Strategy
MARCH
AI: Shaping a Future New Zealand
APRIL
AI Sector Deal
APRIL
Communication on AI for Europe
MAY
AI R&D Strategy
MAY
National Approach for AI
MAY
AI in the Swedish Business and Society
MAY
The Declaration on AI in the Nordic-Baltic Region
JUNE
Common Vision for the Future of AI
JUNE
National Strategy for AI #AIforAll
JUNE
Towards an AI Strategy in Mexico: Harnessing the AI Revolution
SEPTEMBER
Digital Switzerland Strategy
NOVEMBER
AI Strategy
NOVEMBER
White Paper: Shaping the Future of Austria with Robotics and AI
NOVEMBER
White Paper: Malta – Towards an AI Strategy
2019
JANUARY
National AI Strategy for Qatar
FEBRUARY
American AI Initiative
FEBRUARY
AI Portugal 2030
MARCH
AI 4 Belgium Strategy
MARCH
RDI Strategy in AI
MARCH
National Strategy for AI
APRIL
Lithuanian AI Strategy: A Vision for the Future
APRIL
Draft AI Strategy for the Digital Government
APRIL
UAE AI Strategy 2031
APRIL
AI: Australia's Ethics Framework
APRIL
Social Principles for Human-Centric AI
MAY
National AI Strategy of the Czech Republic
MAY
OECD Principles on AI
MAY
AI: A Strategic Vision for Luxembourg
JUNE
AI Mission Austria 2030 (AIM)
JUNE
Leading the Way into the Age of AI
JUNE
The National AI Research and Development Strategic Plan: 2019 Update
JUNE
AI Policy for Sri Lanka
JUNE
UN System-wide Strategic Approach and Roadmap for Supporting Capacity Development on AI
AUGUST
Draft National Strategy on AI
OCTOBER
Strategic Action Plan for AI
OCTOBER
Malta the Ultimate Launchpad: A Strategy and Vision for Artificial Intelligence in Malta 2030
OCTOBER
National AI Strategy 2019 – 2021
OCTOBER
National Strategy: AI Actions for the Use and Development of AI in Cyprus
OCTOBER
National Strategy for the Development of AI by 2030
NOVEMBER
National AI Strategy
NOVEMBER
White Paper: Recommendations for an AI Strategy in Switzerland
NOVEMBER
AI: Solving Problems, Growing the Economy and Improving Quality of Life
NOVEMBER
Draft National Policy for Digital Transformation and AI
2020
JANUARY
The National Strategy for AI
JANUARY
Strategy for the Development of AI in the Republic of Serbia for the Period 2020-2025
FEBRUARY
White Paper On AI – A European approach to excellence and trust
United Nations Interregional Crime and Justice Research Institute
unicri
CENTRE FOR ARTIFICIAL INTELLIGENCE AND ROBOTICS

Source (Wikipedia)

Introduction

The regulation of artificial intelligence is the development of public sector policies and laws for promoting and regulating artificial intelligence (AI); it is therefore related to the broader regulation of algorithms. The regulatory and policy landscape for AI is an emerging issue in jurisdictions globally, including in the European Union. Regulation is considered necessary to both encourage AI and manage associated risks. Regulation of AI through mechanisms such as review boards can also be seen as social means to approach the AI control problem.

Global guidance

The development of a global governance board to regulate AI development was suggested at least as early as 2017. In December 2018, Canada and France announced plans for a G7-backed International Panel on Artificial Intelligence, modeled on the International Panel on Climate Change, to study the global effects of AI on people and economies and to steer AI development. In 2019 the Panel was renamed the Global Partnership on AI, but it is yet to be endorsed by the United States.

The OECD Recommendations on AI were adopted in May 2019, and the G20 AI Principles in June 2019. In September 2019 the World Economic Forum issued ten 'AI Government Procurement Guidelines'. In February 2020, the European Union published its draft strategy paper for promoting and regulating AI.

Regional and national regulation

The regulatory and policy landscape for AI is an emerging issue in jurisdictions globally, including in the European Union. Since early 2016, many national, regional and international authorities have begun adopting strategies, actions plan and policy papers on AI. These documents cover a wide range of topics such as regulation and governance, as well as industrial strategy, research, talent and infrastructure.

China

The regulation of AI in China is mainly governed by the State Council of the PRC's July 8, 2017 "A Next Generation Artificial Intelligence Development Plan" (State Council Document No. 35), in which the Central Committee of the Communist Party of China and the State Council of the People's Republic of China urged the governing bodies of China to promote the development of AI. Regulation of the issues of ethical and legal support for the development of AI is nascent, but policy ensures state control of Chinese companies and over valuable data, including storage of data on Chinese users within the country and the mandatory use of People's Republic of China's national standards for AI, including over big data, cloud computing, and industrial software.

More info: https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/china

European Union

The European Union (EU) is guided by a European Strategy on Artificial Intelligence, supported by a High-Level Expert Group on Artificial Intelligence. In April 2019, the European Commission published its Ethics Guidelines for Trustworthy Artificial Intelligence (AI), following this with its Policy and investment recommendations for trustworthy Artificial Intelligence in June 2019.

On February 2, 2020, the European Commission published its White Paper on Artificial Intelligence - A European approach to excellence and trust. The White Paper consists of two main building blocks, an ‘ecosystem of excellence’ and a ‘ecosystem of trust’. The latter outlines the EU's approach for a regulatory framework for AI. In its proposed approach, the Commission differentiates between 'high-risk' and 'non-high-risk' AI applications. Only the former should be in the scope of a future EU regulatory framework. Whether this would be the case could in principle be determined by two cumulative criteria, concerning critical sectors and critical use. Following key requirements are considered for high-risk AI applications: requirements for training data; data and record-keeping; informational duties; requirements for robustness and accuracy; human oversight; and specific requirements for specific AI applications, such as those used for purposes of remote biometric identification. AI applications that do not qualify as ‘high-risk’ could be governed by voluntary labeling scheme. As regards compliance and enforcement, the Commission considers prior conformity assessments which could include 'procedures for testing, inspection or certification' and/or 'checks of the algorithms and of the data sets used in the development phase'. A European governance structure on AI in the form of a framework for cooperation of national competent authorities could facilitate the implementation of the regulatory framework.

More info: https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

United Kingdom

The UK supported the application and development of AI in business via the Digital Economy Strategy 2015-2018, introduced at the beginning of 2015 by Innovate UK as part of the UK Digital Strategy. In the public sector, guidance has been provided by the Department for Digital, Culture, Media and Sport, on data ethics and the Alan Turing Institute, on responsible design and implementation of AI systems. In terms of cyber security, the National Cyber Security Centre has issued guidance on 'Intelligent Security Tools'.

More info: https://www.gov.uk/government/publications/digital-economy-strategy-2015-2018

United States

Discussions on regulation of AI in the United States have included topics such as the timeliness of regulating AI, the nature of the federal regulatory framework to govern and promote AI, including what agency should lead, the regulatory and governing powers of that agency, and how to update regulations in the face of rapidly changing technology, as well as the roles of state governments and courts.

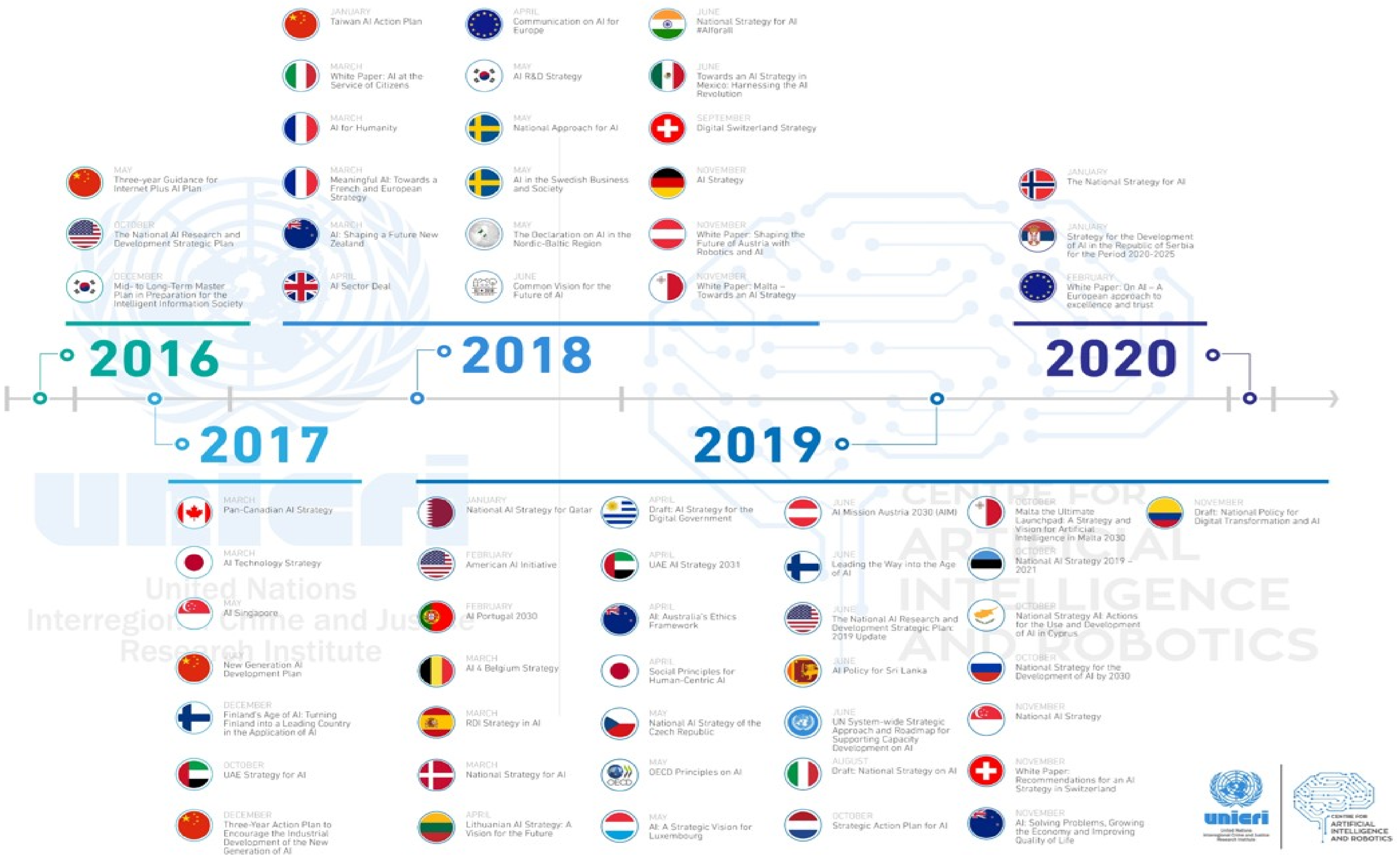
As early as 2016, the Obama administration had begun to focus on the risks and regulations for artificial intelligence. In a report titled Preparing For the Future of Artificial Intelligence, the National Science and Technology Council set a precedent to allow researchers to continue to develop new AI technologies with few restrictions. It is stated within the report that "the approach to regulation of AI-enabled products to protect public safety should be informed by assessment of the aspects of risk....". These risks would be the principle reason to create any form of regulation, granted that any existing regulation would not apply to AI technology.

The first main report was the National Strategic Research and Development Plan for Artificial Intelligence. On August 13, 2018, Section 1051 of the Fiscal Year 2019 John S. McCain National Defense Authorization Act (P.L. 115-232) established the National Security Commission on Artificial Intelligence "to consider the methods and means necessary to advance the development of artificial intelligence, machine learning, and associated technologies to comprehensively address the national security and defense needs of the United States." Steering on regulating security-related AI is provided by the National Security Commission on Artificial Intelligence. The Artificial Intelligence Initiative

Act (S.1558) is a proposed bill that would establish a federal initiative designed to accelerate research and development on AI for, inter alia, the economic and national security of the United States.

On January 7, 2019, following an Executive Order on Maintaining American Leadership in Artificial Intelligence, the White House's Office of Science and Technology Policy released a draft Guidance for Regulation of Artificial Intelligence Applications, which includes ten principles for United States agencies when deciding whether and how to regulate AI. In response, the National Institute of Standards and Technology has released a position paper, the National Security Commission on Artificial Intelligence has published an interim report, and the Defense Innovation Board has issued recommendations on the ethical use of AI. A year later, the administration called for comments on further deregulation in another draft of its Guidance for Regulation of Artificial Intelligence Applications.

More info:
https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence



2016

MAY
Three-year Guidance for Internet Plus AI Plan

OCTOBER
The National AI Research and Development Strategic Plan

DECEMBER
Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society

2017

MARCH
Pan-Canadian AI Strategy

MARCH
AI Technology Strategy

MAY
AI Singapore

MAY
New Generation AI Development Plan

DECEMBER
Finland's Age of AI: Turning Finland into a Leading Country in the Application of AI

OCTOBER
UAE Strategy for AI

DECEMBER
Three-Year Action Plan to Encourage the Industrial Development of the New Generation of AI

2018

JANUARY
Taiwan AI Action Plan

MARCH
White Paper: AI at the Service of Citizens

MARCH
AI for Humanity

MARCH
Meaningful AI: Towards a French and European Strategy

MARCH
AI: Shaping a Future New Zealand

APRIL
AI Sector Deal

APRIL
Communication on AI for Europe

MAY
AI R&D Strategy

MAY
National Approach for AI

MAY
AI in the Swedish Business and Society

MAY
The Declaration on AI in the Nordic-Baltic Region

JUNE
Common Vision for the Future of AI

JUNE
National Strategy for AI #AIforall

JUNE
Towards an AI Strategy in Mexico: Harnessing the AI Revolution

SEPTEMBER
Digital Switzerland Strategy

NOVEMBER
AI Strategy

NOVEMBER
White Paper: Shaping the Future of Austria with Robotics and AI

NOVEMBER
White Paper: Malta – Towards an AI Strategy

2019

JANUARY
National AI Strategy for Qatar

FEBRUARY
American AI Initiative

FEBRUARY
AI Portugal 2030

MARCH
AI 4 Belgium Strategy

MARCH
RDI Strategy in AI

MARCH
National Strategy for AI

APRIL
Lithuanian AI Strategy: A Vision for the Future

APRIL
Draft: AI Strategy for the Digital Government

APRIL
UAE AI Strategy 2031

APRIL
AI: Australia's Ethics Framework

APRIL
Social Principles for Human-Centric AI

MAY
National AI Strategy of the Czech Republic

MAY
OECD Principles on AI

MAY
AI: A Strategic Vision for Luxembourg

JUNE
AI Mission Austria 2030 (AIM)

JUNE
Leading the Way into the Age of AI

JUNE
The National AI Research and Development Strategic Plan: 2019 Update

JUNE
AI Policy for Sri Lanka

JUNE
UN System-wide Strategic Approach and Roadmap for Supporting Capacity Development on AI

AUGUST
Draft: National Strategy on AI

OCTOBER
Strategic Action Plan for AI

OCTOBER
Malta the Ultimate Launchpad: A Strategy and Vision for Artificial Intelligence in Malta 2030

OCTOBER
National AI Strategy 2019 – 2021

OCTOBER
National Strategy AI: Actions for the Use and Development of AI in Cyprus

OCTOBER
National Strategy for the Development of AI by 2030

NOVEMBER
National AI Strategy

NOVEMBER
White Paper: Recommendations for an AI Strategy in Switzerland

NOVEMBER
AI: Solving Problems, Growing the Economy and Improving Quality of Life

NOVEMBER
Draft: National Policy for Digital Transformation and AI

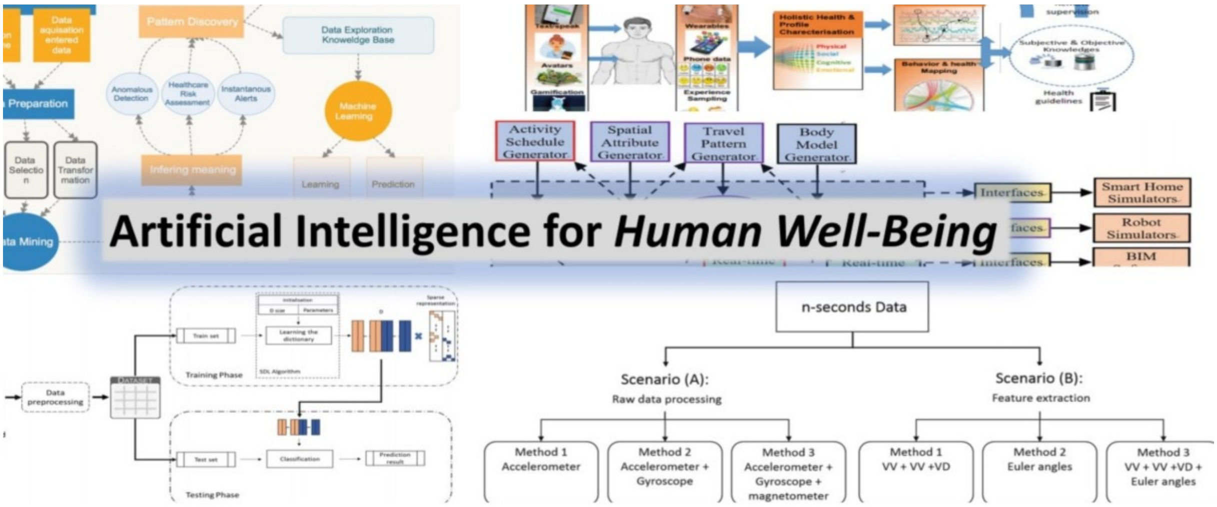
2020

JANUARY
The National Strategy for AI

JANUARY
Strategy for the Development of AI in the Republic of Serbia for the Period 2020-2025

FEBRUARY
White Paper: On AI – A European approach to excellence and trust

A Collection of Papers & Articles on AI for Human Well-Being



Numerous AI initiatives are underway in the health sector. Some of these are aimed at promoting mental health and well-being. In the following, I would like to present some promising approaches/articles from this area.

Improving Access and Mental Health for Youth Through Virtual Models of Care

"The overall objective of this research is to evaluate the use of a mobile health smartphone application (app) to improve the mental health of youth between the ages of 14–25 years, with symptoms of anxiety/depression. This project includes 115 youth who are accessing outpatient mental health services at one of three hospitals and two community agencies. The youth and care providers are using eHealth technology to enhance care. The technology uses mobile questionnaires to help promote self-assessment and track changes to support the plan of care. The technology also allows secure virtual treatment visits that youth can participate in through mobile devices. This longitudinal study uses participatory action research with mixed methods. The majority of participants identified themselves as Caucasian (66.9%). Expectedly, the demographics revealed that Anxiety Disorders and Mood Disorders were highly prevalent within the sample (71.9% and 67.5% respectively). Findings from the qualitative summary established that both staff and youth found the software and platform beneficial."

Source: **Improving Access and Mental Health for Youth Through Virtual Models of Care**

Combining Human and Artificial Intelligence for Analyzing Health Data

"Artificial intelligence (AI) systems are increasingly capable of analyzing health data such as medical images (e.g., skin lesions) and test results (e.g., ECGs). However, because it can be difficult to determine when an AI-generated diagnosis should be trusted and acted upon—especially when it conflicts with a human-generated one—many AI systems are not utilized effectively, if at all. Similarly, advances in information technology have made it possible to quickly solicit multiple diagnoses from diverse groups of people throughout the world, but these technologies are underutilized because it is difficult to determine which of multiple diagnoses should be trusted and acted upon. Here, I propose a method of soliciting and combining multiple diagnoses that will harness the collective intelligence of both human and artificial intelligence for analyzing health data. "

source: **Combining Human and Artificial Intelligence for Analyzing Health Data**

Design of a Framework for Wellness Determination and Subsequent Recommendation with Personal Informatics

"Due to the advances in medical science, increasing health consciousness, improved quality of food, the average human life span has increased to a great extent. On the other hand, stresses of modern life, overwork and less sleep, increased usage of digital devices and internet, less exercise, are leading us to poor quality of life. Elderly people are more vulnerable to reduced life quality due to deterioration of both physical and mental health. People at any age need to maintain a minimum level of wellbeing to pursue his or her daily activities to lead a fulfilling life. Thus the need of assessing and restoring wellness is very important. Fortunately the progress of information and communication technologies provide use sensor devices and computing platform to feel, monitor and restore the wellness."

source: **Design of a Framework for Wellness Determination and Subsequent Recommendation with Personal Informatics**

Combined Machine Learning and Semantic Modelling for Situation Awareness and Healthcare Decision Support

"The average of global life expectancy at birth was 72 years in 2016 , however, the global healthy life expectancy at birth was only 63.3 years in the same year, 2016. Living a long life is not any more as challenging as assuring active and associated life. We propose in this paper an IoT based holistic remote health monitoring system for chronically ill and elderly patients. It supports smart clinical decision help and prediction. The patient heterogeneous vital signs and contexts gathered from wore and surrounding sensors are semantically simplified and modeled via a validated ontology composed by FOAF (Friend of a Friend), SSN (Semantic Sensors Network)/SOSA (Sensor, Observation, Sample and Actuator) and ICNP (International Classification Nursing Practices) ontologies. The reasoner engine is based on a scalable set of inference rules cohesively integrated with a ML (Machine Learning) algorithm to ensure predictive analytic and preventive personalized health services. Experimental results prove the efficiency of the proposed system."

source: **Combined Machine Learning and Semantic Modelling for Situation Awareness and Healthcare Decision Support**

Automatic Daily Activity Schedule Planning for Simulating Smart House with Elderly People Living Alone

"A simulation tool that supports developers to build scenarios automatically in multiple simulation platforms is proposed. As an essential part of this simulator, this study proposed an activity schedule generator to mimic the daily life of elderly people living alone. This generator outperforms existing methods of activity schedule planning in three aspects: 1) it is adaptive to the layout of a simulated smart house; 2) there is no unspecified time in the timeline of generated schedules; and 3) it generates stable, but not tedious schedules for a number of days. A real-time location data generator is proposed to convert generated schedules to simulated real-time location data of the resident, and a proposed interface converts these simulated location data to simulated records of virtual passive infrared (PIR) sensors,"

source: **Automatic Daily Activity Schedule Planning for Simulating Smart House with Elderly People Living Alone**

A Novel On-Wrist Fall Detection System Using Supervised Dictionary Learning Technique

"Wrist-based fall detection system provides a very comfortable and multi-modal healthcare solution, especially for elderly risking falls. However, the wrist location presents a very challenging and unstable spot to distinguish falls among other daily activities. In this paper, we propose a Supervised Dictionary Learning approach for wrist-based fall detection. Three Dictionary learning algorithms for classification are invoked in this study, namely SRC, FDDL, and LRSDL. To extract the best descriptive representation of the signal data we followed different preprocessing scenarios based on accelerometer, gyroscope, and magnetometer. A considerable overall performance was obtained by the SRC algorithms reaching respectively 99.8%, 100%, and 96.6% of accuracy, sensitivity, and specificity using raw data provided by a triaxial accelerometer, according"

source: **A Novel On-Wrist Fall Detection System Using Supervised Dictionary Learning Technique**

Mindful Technologies Research and Developments in Science and Art

"This paper outlines three projects that lay the foundation for a trans-disciplinary approach to the creation of interactive, multi-sensory devices combining biofeedback, virtual reality, and physical/virtual human-machine interactions. We explore new possibilities for interoperability and enhancing interoception and mindfulness with potential research contributions for novel personal, professional and medical applications."

source: **Mindful Technologies Research and Developments in Science and Art**

Prioritizing Human Well-being in the Age of Artificial Intelligence (IEEE)

“The idea is that the changes that are brought forward by these digital technologies could be monitored across all the various dimensions of human well-being.”

Fabrice Murtin, Senior Economist , Household Statistics and progress Measurement Division of the OECD Statistics Directorate

“One of the first aims of the EU is the promotion of its values and the well-being of its people.”

Salla Saastamoinen – Director, Directorate A, Civil and Commercial Justice, Directorate -General (DG) for Panel One, from left to right: Salla Saastamoinen, Justice and Consumers (JUST), European Commission

"On the various recommendations raised you can ask questions about how to address the well-being of the user to come up with concrete solutions. This is the game changer for how we rethink development."

Raja Chatila - Chair of The IEEE Global Initiative

"What companies are realizing is that well-being and human value are the central motivation for their innovation."

Virginia Dignum - Associate Professor, Delft University of Technology

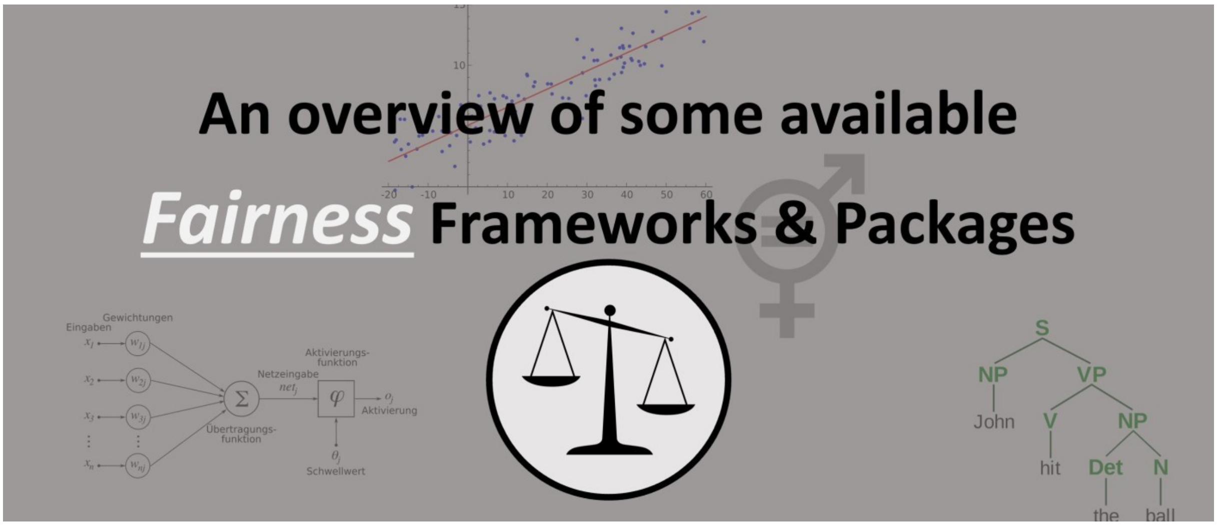
source: **IEEE**

This might be also of interest:

- **A collection of recommendable papers and articles on AI-Governance**
- **A collection of recommendable papers and articles on Explainable AI (XAI)**
- **Data Ethics: 7 Points to Consider**

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

An overview of some available Fairness Frameworks & Packages



Content

1. The LinkedIn Fairness Toolkit (LiFT)

The LinkedIn Fairness Toolkit (LiFT) is a Scala/Spark library that enables the measurement of fairness in large scale machine learning workflows. The library can be deployed in training and scoring workflows to measure biases in training data, evaluate fairness metrics for ML models, and detect statistically significant differences in their performance across different subgroups. It can also be used for ad-hoc fairness analysis.

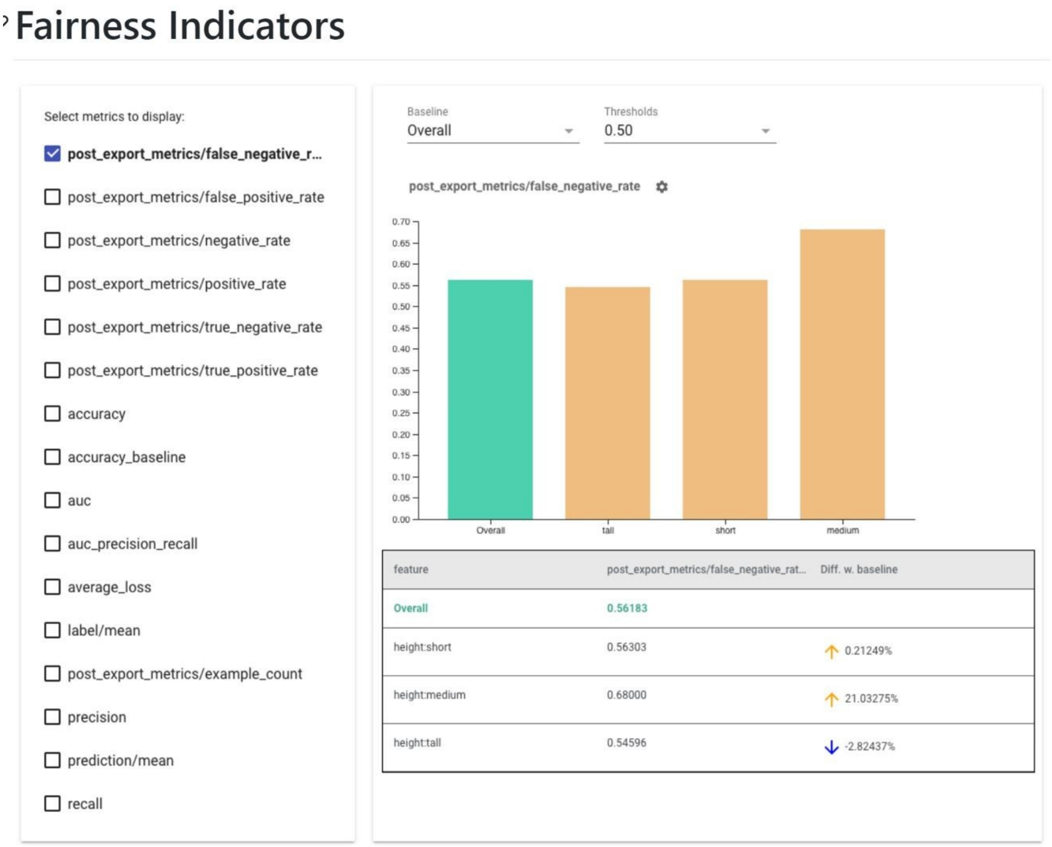
This library was created by **Sriram Vasudevan** and **Krishnaram Kenthapadi** (work done while at LinkedIn).

More info: **https://github.com/linkedin/LiFT**

2. Fairness-indicators: Tensorflow's Fairness Evaluation and Visualization Toolkit (Google)

Fairness Indicators is designed to support teams in evaluating, improving, and comparing models for fairness concerns in partnership with the broader Tensorflow toolkit.

The tool is currently actively used internally by many of our products. We would love to partner with you to understand where Fairness Indicators is most useful, and where added functionality would be valuable. Please reach out at **tfx@tensorflow.org**.



Fairness Indicators

Select metrics to display:

- ☑ post_export_metrics/false_negative_r...
- ☐ post_export_metrics/false_positive_rate
- ☐ post_export_metrics/negative_rate
- ☐ post_export_metrics/positive_rate
- ☐ post_export_metrics/true_negative_rate
- ☐ post_export_metrics/true_positive_rate
- ☐ accuracy
- ☐ accuracy_baseline
- ☐ auc
- ☐ auc_precision_recall
- ☐ average_loss
- ☐ label/mean
- ☐ post_export_metrics/example_count
- ☐ precision
- ☐ prediction/mean
- ☐ recall

Baseline: Overall

Thresholds: 0.50

post_export_metrics/false_negative_rate

feature	post_export_metrics/false_negative_rat...	Diff. w. baseline
Overall	0.56183	
height:short	0.56303	↑ 0.21249%
height:medium	0.68000	↑ 21.03275%
height:tall	0.54596	↓ -2.82437%

More info: **https://github.com/tensorflow/fairness-indicators**

3. AI Fairness 360 (IBM)

The AI Fairness 360 toolkit is an extensible open-source library containg techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle. AI Fairness 360 package is available in both Python and R.

The AI Fairness 360 package includes

- a comprehensive set of metrics for datasets and models to test for biases,

- explanations for these metrics, and

- algorithms to mitigate bias in datasets and models. It is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

More info: **https://github.com/Trusted-AI/AIF360**

4. Fairlearn: Fairness in machine learning mitigation algorithms (Microsoft)

Fairlearn is a Python package that empowers developers of artificial intelligence (AI) systems to assess their system's fairness and mitigate any observed unfairness issues. Fairlearn contains mitigation algorithms as well as a Jupyter widget for model assessment. Besides the source code, this repository also contains Jupyter notebooks with examples of Fairlearn usage.

More Info: **https://github.com/fairlearn/fairlearn**

5. Algofairness

BlackBoxAuditing

This repository contains a sample implementation of Gradient Feature Auditing (GFA) meant to be generalizable to most datasets. For more information on the repair process, see our paper on Certifying and Removing Disparate Impact. For information on the full auditing process, see our paper on Auditing Black-box Models for Indirect Influence.

More info: **https://github.com/algofairness/BlackBoxAuditing**

fairness-comparison

This repository is meant to facilitate the benchmarking of fairness aware machine learning algorithms.

The associated paper is:

A comparative study of fairness-enhancing interventions in machine learning by Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. **https://arxiv.org/abs/1802.04422**

More info: **https://github.com/algofairness/fairness-comparison**

fatconference-2019-toolkit-tutorial

More info: **https://github.com/algofairness/fatconference-2019-toolkit-tutorial**

fatconference-2018-auditing-tutorial

More info: **https://github.com/algofairness/fatconference-2018-auditing-tutorial**

runaway-feedback-loops-src

More info: **https://github.com/algofairness/runaway-feedback-loops-src**

Knight

More info: **https://github.com/algofairness/knight**

6. FairSight: Visual Analytics for Fairness in Decision Making

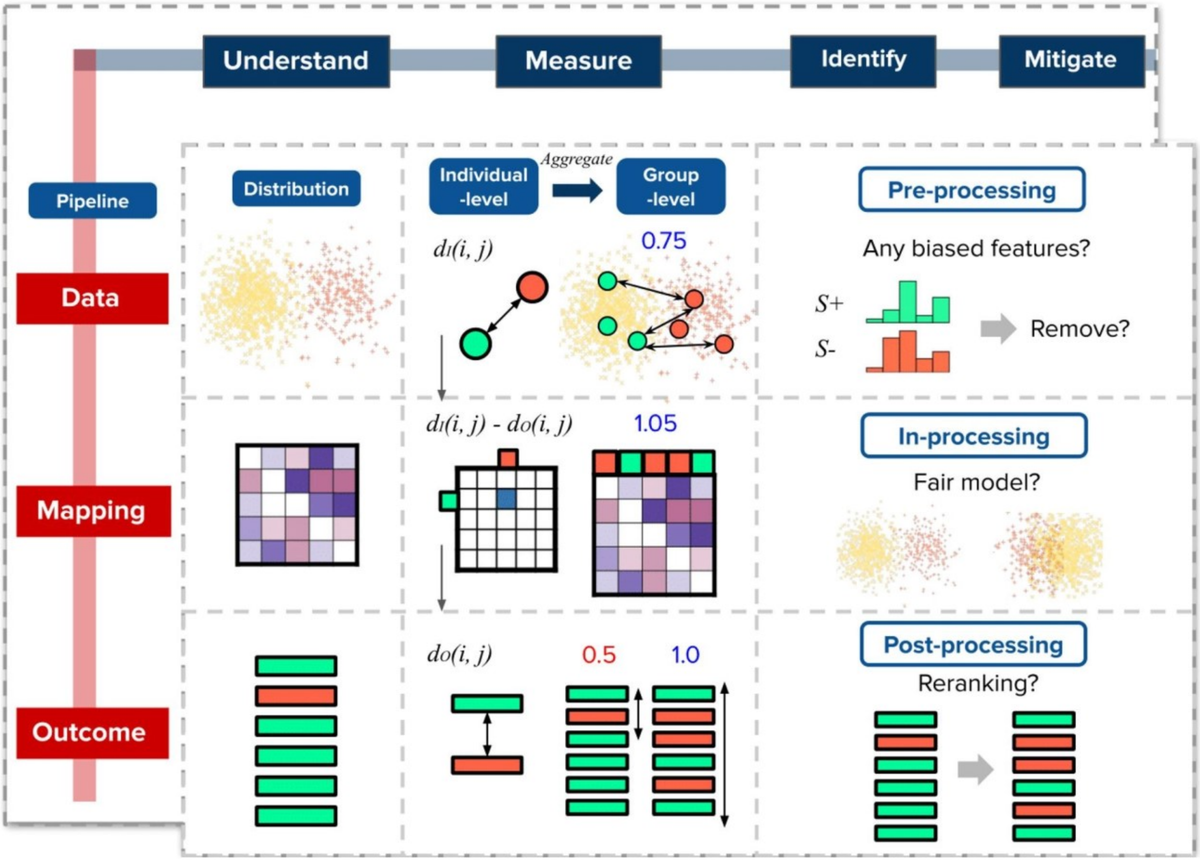
FairSight is a viable fair decision-making system to assist decision makers in achieving fair decision making through the machine learning workflow.

Specification

React: Frontend framework for rendering and communicating with data django: Python-based backend framework for serving API of data and running machine learning work

scss: The stylesheet grammar for more flexible structure

d3.js: Javascript-based visualization library



More info: https://github.com/ayong8/FairSight

7. Aequitas: Bias and Fairness Audit Toolkit

Aequitas is an open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools.

More info: https://github.com/dssg/aequitas

8. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models

Concerns within the machine learning community and external pressures from regulators over the vulnerabilities of machine learning algorithms have spurred on the fields of explainability, robustness, and fairness. Often, issues in explainability, robustness, and fairness are confined to their specific sub-fields and few tools exist for model developers to use to simultaneously build their modeling pipelines in a transparent, accountable, and fair way. This can lead to a bottleneck on the model developer's side as they must juggle multiple methods to evaluate their algorithms. In this paper, we present a single framework for analyzing the robustness, fairness, and explainability of a classifier. The framework, which is based on the generation of counterfactual explanations1 through a custom genetic algorithm, is flexible, model-agnostic, and does not require access to model internals. The framework allows the user to calculate robustness and fairness scores for individual models and generate explanations for individual predictions which provide a means for actionable recourse (changes to an input to help get a desired outcome). This is the first time that a unified tool has been developed to address three key issues pertaining towards building a responsible artificial intelligence system.

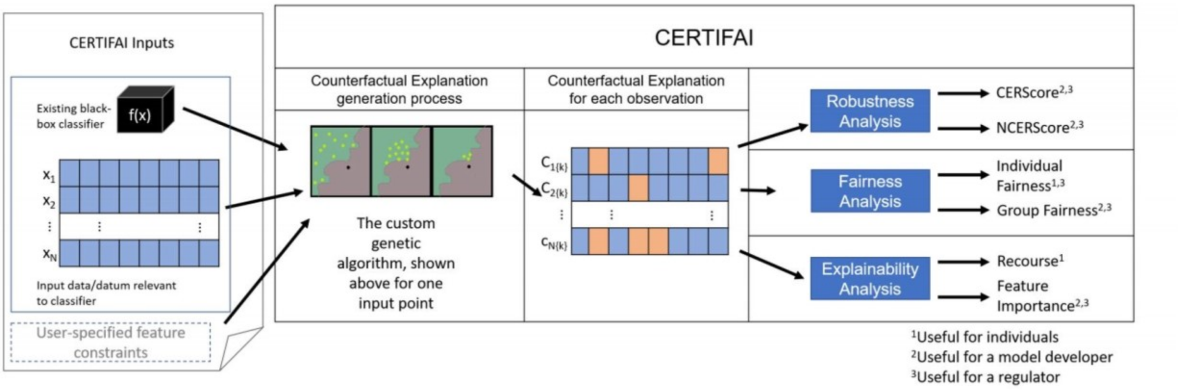

CERTIFAI Inputs
Existing black-box classifier
f(x)
x_1
x_2
x_N
Input data/datum relevant to classifier
User-specified feature constraints
CERTIFAI
Counterfactual Explanation generation process
The custom genetic algorithm, shown above for one input point
Counterfactual Explanation for each observation
$C_{1(k)}$
$C_{2(k)}$
$C_{N(k)}$
Robustness Analysis
CERScore[2,3]
NCERScore[2,3]
Fairness Analysis
Individual Fairness[1,3]
Group Fairness[2,3]
Explainability Analysis
Recourse[1]
Feature Importance[2,3]
[1]Useful for individuals
[2]Useful for a model developer
[3]Useful for a regulator

Figure 2: The CERTIFAI framework. Given a black-box ML model and input data along with optional user-specified feature constraints (such as feature type, range, etc.), the method generates counterfactual explanations using a genetic algorithm. The explanations can then be used for three purposes: explainability, fairness and robustness. k represents the number of explanations per input which can be set by the user for recourse purposes and is set to 1 for the feature importance, fairness and robustness analysis. On the right, we show how each of CERTIFAI's attributes is useful for different stakeholders using the tool

More info: CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models

9. ML-fairness-gym: Google's implementation based on OpenAI's Gym

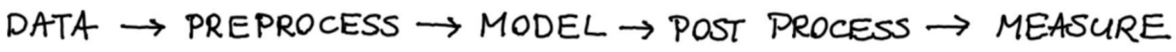
ML-fairness-gym is a set of components for building simple simulations that explore the potential long-run impacts of deploying machine learning-based decision systems in social environments. As the importance of machine learning fairness has become increasingly apparent, recent research has focused on potentially surprising long term behaviors of enforcing measures of fairness that were originally defined in a static setting. Key findings have shown that under specific assumptions in simplified dynamic simulations, long term effects may in fact counteract the desired goals.

More info: **https://github.com/google/ml-fairness-gym**

10. scikit-fairness

The goal of this project is to attempt to consolidate fairness related metrics, transformers and models into a package that (hopefully) will become a contribution project to scikit-learn.

Fairness, in data science, is a complex unsolved problem for which many tactics are proposed - each with their own advantage and disadvantages. This packages aims to make these tactics readily available, therefore enabling users to try and evaluate different fairness techniques.



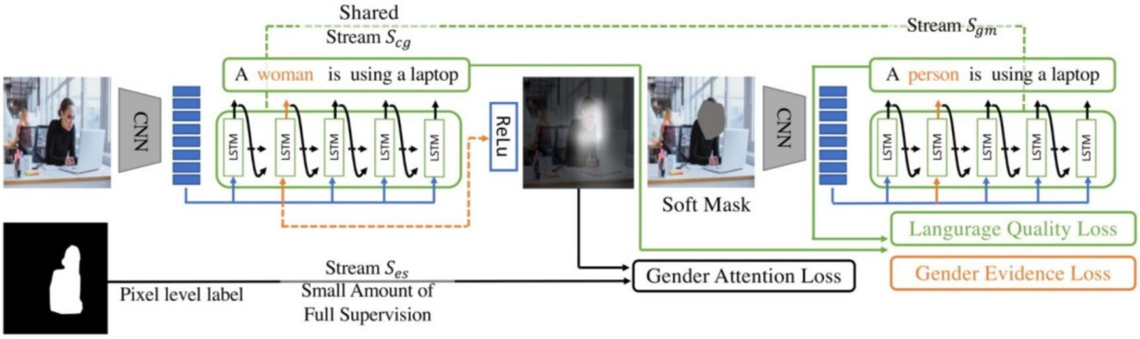
More info: **https://github.com/koaning/scikit-fairness**

11. Mitigating Gender Bias In Captioning System

This is the pytorch implemention for paper “Mitigating Gender Bias In Captioning system”. Recent studies have shown that captioning datasets, such as the COCO dataset, may contain severe social bias which could potentially lead to unintentional discrimination in learning models. In this work, we specifically focus on the gender bias problem.

Image Captioning Model with Guided Attention

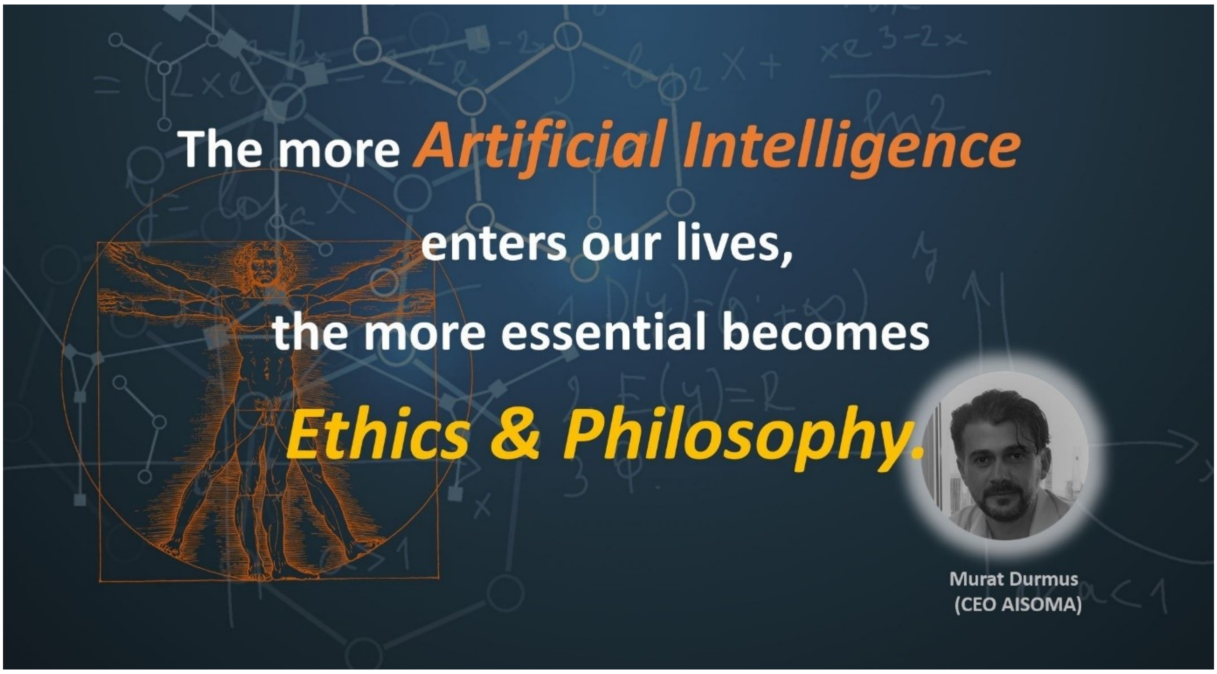
We propose a novel Guided Attention Image Captioning model (GAIC) to mitigate gender bias by self-supervising on model’s visual attention. GAIC has two complementary streams to encourage the model to explore correct gender features. The training pipeline can seamlessly add extra supervision to accelerate the self-exploration process. Besides, GAIC is model-agnostic and can be easily applied to various captioning models.



More info:

https://github.com/CaptionGenderBias2020/Mitigating_Gender_Bias_In_Captioning_System

This might be also of interest: A collection of useful Slides & Quotes on AI-Ethics and XAIt



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