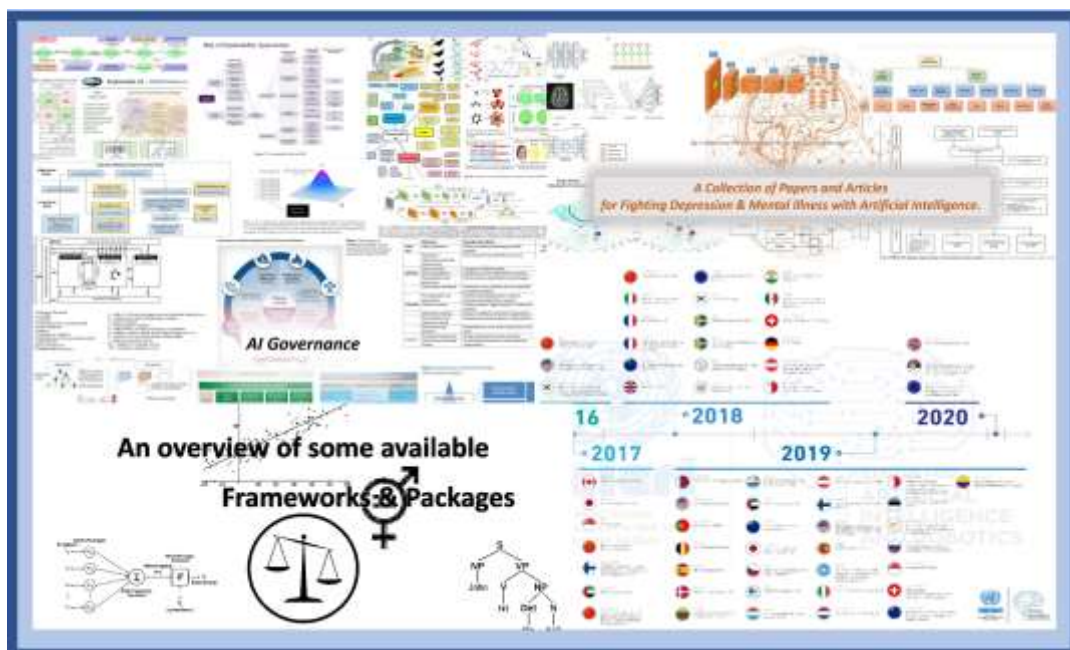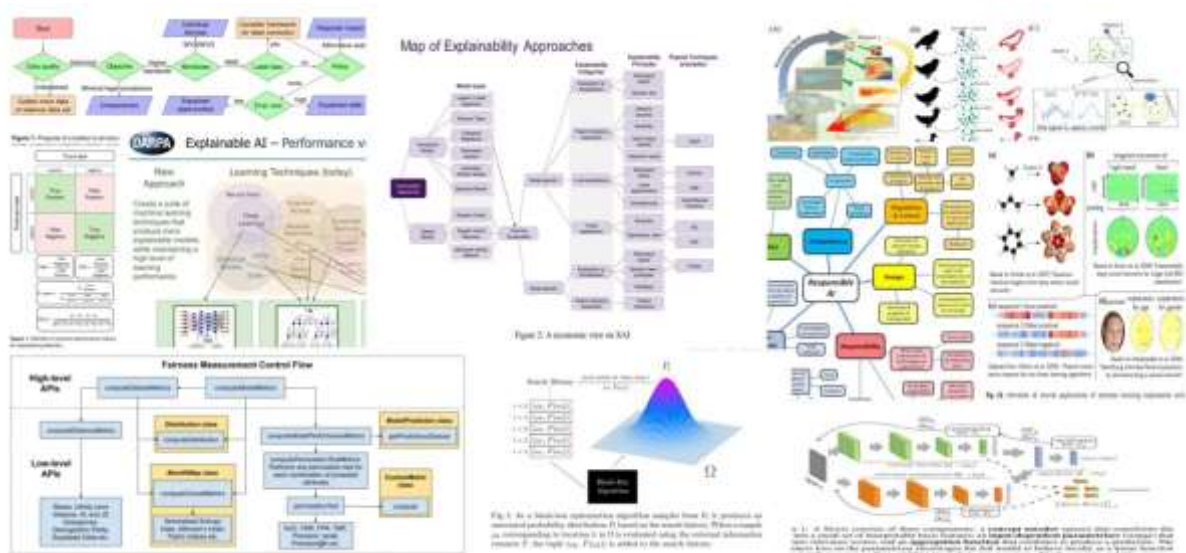# The Full Compilation
A Collection of Recommendable Papers & Articles on AI, XAI, AI-Ethics, AI-Governance, Regulation, AIEd AI for Mental-Health, Well-Being and Philosophy

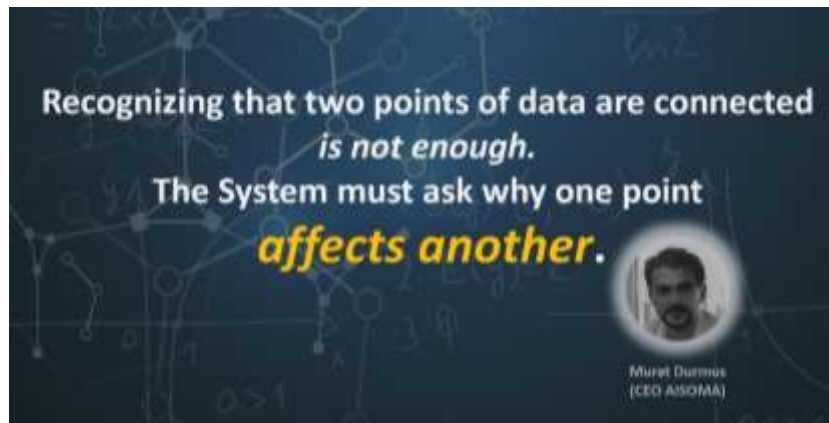# A collection of recommendable papers and articles on Explainable AI (XAI)

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*)

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**

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.

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 depression[1–3], suicidality[4–6], eating disorders[7,8], and schizophrenia[9], among others with high accuracy (80–90%). In addition to mental disorders, these approaches are starting to assess related symptomatology, such as self-harm[8], stress[10], and the severity of mental illness[11] 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 Facebook[12]. 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*
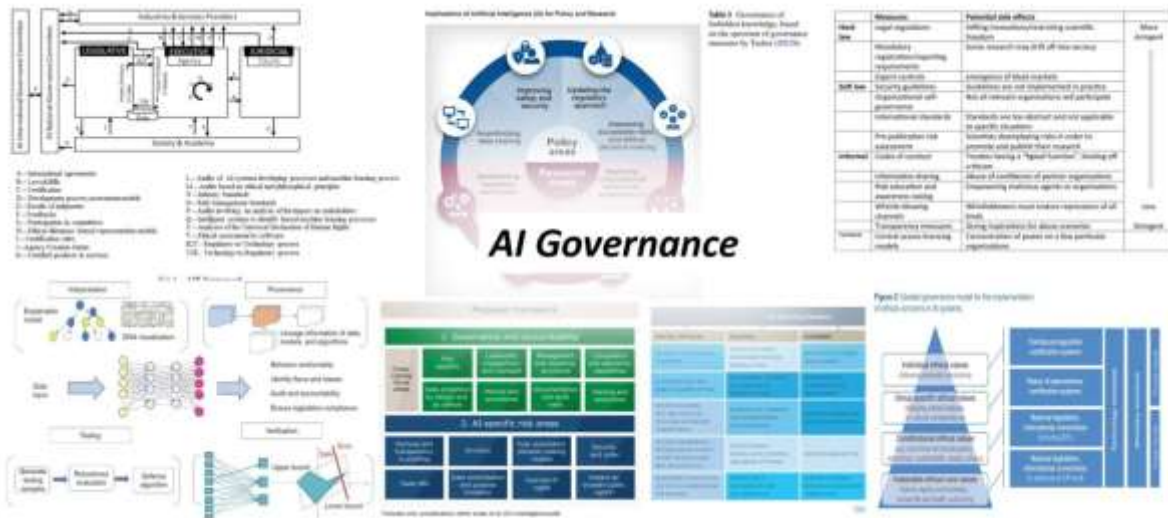
The development and use of artificial intelligence (AI) technologies can bring about many benefits – from transforming businesses and improving labour productivity to enhancing quality of life. However, ***there is an increasing concern regarding the risk of harm associated with the use of AI technologies 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

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

2017

2018

2019

2020

2016

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

OCTOBER — The National AI Research and Development Strategic Plan

MAY — Three-Year Guidance for Internet Plus AI Plan

2017

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

OCTOBER — UAE Strategy for AI

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

JULY — New Generation AI Development Plan

MAY — AI Singapore

MARCH — AI Technology Strategy

MARCH — Pan-Canadian AI Strategy

OCTOBER — AI Sector Deal

MAY — AI: Shaping a Future New Zealand

MARCH — Meaningful AI: Towards a French and European Strategy

MARCH — AI for Humanity

MARCH — White Paper: AI at the Service of Citizens

JANUARY — Taiwan AI Action Plan

2018

APRIL — Lithuanian AI Strategy: A Vision for the Future

MARCH — National Strategy for AI

MARCH — RDI Strategy in AI

MARCH — AI 4 Belgium Strategy

FEBRUARY — AI Portugal 2030

FEBRUARY — American AI Initiative

JANUARY — National AI Strategy for Qatar

JUNE — AI a Common Vision for the Future of AI

MAY — The Declaration on AI in the Nordic-Baltic Region

MAY — AI in the Swedish Business and Society

MAY — National Approach for AI

MAY — AI R&D Strategy

APRIL — Communication on AI for Europe

2019

MAY — AI A Strategic Vision for Luxembourg

MAY — OECD Principles on AI

MAY — National AI Strategy of the Czech Republic

APRIL — Social Principles for Human-Centric AI

APRIL — AI Australia's Ethics Framework

APRIL — UAE AI Strategy 2031

APRIL — Draft: AI Strategy for the Digital Government

NOVEMBER — White Paper: Malta — Towards an AI Strategy

NOVEMBER — White Paper: Future of Austria with Robotics and AI

NOVEMBER — AI Strategy

NOVEMBER — Towards an AI Strategy in Mexico: Harnessing the AI Revolution

SEPTEMBER — Digital Switzerland Strategy

JUNE — National Strategy for AI #AIforall

OCTOBER — Strategic Action Plan for AI

OCTOBER — Draft National Strategy on AI

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

JUNE — National Strategy and Development of AI in Cyprus

MAY — The National AI Research and Development Strategic Plan: 2019 Update

MAY — AI Policy for Sri Lanka

JUNE — Leading the Way into the Age of AI

2030 — AI Mission Austria 2030 (AIM)

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

NOVEMBER — Recommendations for an AI Strategy in Switzerland

NOVEMBER — National AI Strategy

NOVEMBER — National Strategy for the Development of AI by 2030

OCTOBER — National AI Strategy 2019 – 2027

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

SEPTEMBER — Draft National Policy for Digital Transformation and AI

2020

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

FEBRUARY — Strategy for the Development of AI in the Republic of Serbia for the Period 2020–2025

JANUARY — The National Strategy for AI

United Nations Interregional Crime and Justice Research Institute

UNICRI

CENTRE FOR ARTIFICIAL INTELLIGENCE AND ROBOTICS

27

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**

A Collection of Recommendable Papers and Articles on AI in Education (AIEd)
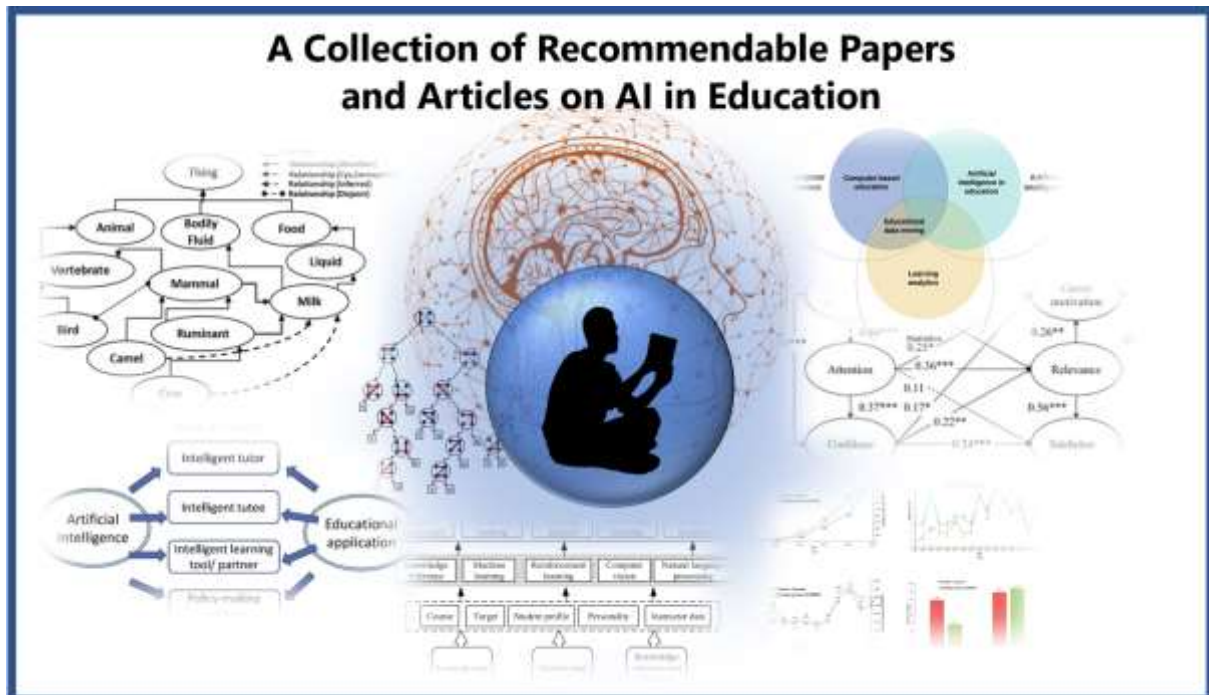
A multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers

Highlights

- Present a multi-perspectives study on Artificial Intelligence in Education (AIEd).
- Analyze relevant funds, conferences, journals, software, institutions, and scholars.
- Demonstrate increasing interest in and impact of AIEd.
- Highlight the significance and necessity to launch specialized AIEd journal.
- Introduce a new Elsevier journal Computers & Education: Artificial Intelligence.

With the rapid development of artificial intelligence (AI) technologies and a continuously growing interest in their application in educational contexts, there has been significant growth in the scientific literature in relation to the application of AI in education (AIEd). This study aims to present multiple perspectives on the development of AIEd in terms of relevant grants, conferences, journals, software

tools, article trends, top issues, institutions, and researchers to provide an overview of AIEd for its further development and implementation

Source: A multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers - ScienceDirect

The Ethical Framework for AI in Education

Great and important work by The Institute for Ethical AI in Education

The topic of AI is still ignored in education (at least in most countries). We have to prepare students for the Age of AI on time. Of course, with a hefty dose of ethics. There is no time to lose. The coming generation will have to deal more with ethical issues than with technological ones. Therefore: Solving the tech industry's ethics issues must start in the classrooms.

"The Ethical Framework for AI in Education Having consulted with a wide range of stakeholders in these ways, the Institute is now in a position to put forward The Ethical Framework for Al In Education. The Framework is grounded in a shared vision of ethical Al in education and will help to enable all learners to benefit optimally from Al in education, whilst also being protected against the risks this technology presents. The Framework is aimed at those making procurement and application decisions relevant to AI in education"

Link (pdf): The-Ethical-Framework-for-AI-in-Education-Institute-for-Ethical-AI-in-Education-Final-Report.pdf (b-cdn.net)

Vision, challenges, roles and research issues of Artificial Intelligence in Education

The rapid advancement of computing technologies has facilitated the implementation of AIED (Artificial Intelligence in Education) applications. AIED refers to the use of AI (Artificial Intelligence) technologies or application programs in educational settings to facilitate teaching, learning, or decision making. With the help of AI technologies, which simulate human intelligence to make inferences, judgments, or predictions, computer systems can provide personalized guidance, supports, or feedback to students as well as assisting teachers or policymakers in making decisions. Although AIED has been identified as the primary research focus in the field of computers and education, the interdisciplinary nature of AIED presents a unique challenge for researchers with different disciplinary backgrounds. In this

paper, we present the definition and roles of AIED studies from the perspective of educational needs. We propose a framework to show the considerations of implementing AIED in different learning and teaching settings. The structure can help guide researchers with both computers and education backgrounds in conducting AIED studies. We outline 10 potential research topics in AIED that are of particular interest to this journal. Finally, we describe the type of articles we like to solicit and the management of the submissions.

Source: Vision, challenges, roles and research issues of Artificial Intelligence in Education - ScienceDirect

Application and theory gaps during the rise of Artificial Intelligence in Education

Highlights

- **Conduct a systematic review of 45 highly cited AIEd articles.**
- **Identify the application and theory gaps during the rise of AI in education.**
- **Establish the linkage between extant AIEd studies and future trends.**
- **Clarify the definitions of AIEd from broad and narrow senses.**

Considering the increasing importance of Artificial Intelligence in Education (AIEd) and the absence of a comprehensive review on it, this research aims to conduct a comprehensive and systematic review of influential AIEd studies. We analyzed 45 articles in terms of annual distribution, leading journals, institutions, countries/regions, the most frequently used terms, as well as theories and technologies adopted. We also evaluated definitions of AIEd from broad and narrow perspectives and clarified the relationship among AIEd, Educational Data Mining, Computer-Based Education, and Learning Analytics. Results indicated that: 1) there was a continuingly increasing interest in and impact of AIEd research; 2) little work had been conducted to bring deep learning technologies into educational contexts; 3) traditional AI technologies, such as natural language processing were commonly adopted in educational contexts, while more advanced techniques were rarely adopted, 4) there was a lack of studies that both employ AI technologies and engage deeply with educational theories.

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

The use of Artificial Intelligence (AI) in education

There are two different types of AI in wide use today. Recent developments have focused on data-driven machine learning, but in the last decades, most AI applications in education (AIEd) have been based on representational / knowledge-based AI.

Data-driven AI uses a programming paradigm that is new to most computing professionals. It requires competences which are different from traditional programming and computational thinking. It opens up new ways to use computing and digital devices. But the development of state-of-the-art AI is now starting to exceed the computational capacity of the largest AI developers. The recent rapid developments in data-driven AI may not be sustainable.

The impact of AI in education will depend on how learning and competence needs change, as AI will be widely used in the society and economy. AIEd should be used to help schools and educational institutions in transforming learning for the future.

Many AIEd systems have been developed over the years, but few of these have shown clear scientific impact on learning. Evidence is lacking partly because the contexts of teaching and learning vary across classrooms, schools, educational systems, and countries. Local knowledge and capacity is critical for effective adoption and shaping of AIEd, and new scaling models are needed. Co-design of AIEd with teachers is a possible way to advance new scaling models.

AI has a great potential in compensating learning difficulties and supporting teachers. The Union/ the EU needs a "clearing house" that helps teachers and policy-makers make sense of the fast developments in this area

Source: The use of Artificial Intelligence (AI) in education – Research4Committees

10 Roles For Artificial Intelligence In Education

For decades, science fiction authors, futurists, and movie makers alike have been predicting the amazing (and sometimes catastrophic) changes that will arise with the advent of widespread artificial intelligence. So far, AI hasn't made any such crazy waves, and in many ways has quietly become ubiquitous in numerous aspects of our daily lives. From the intelligent sensors that help us take perfect pictures, to the automatic parking features in cars, to the sometimes frustrating personal assistants in smartphones, artificial intelligence of one kind of another is all around us, all the time.

Source: 10 Roles For Artificial Intelligence In Education (teachthought.com)

Unleashing the power of AI for education

Educators today are in agreement: They need an AI strategy. But many institutions don't know how to implement one—or where to start.

Source: Unleashing the power of AI for education | MIT Technology Review

Artificial Intelligence in Education: A Review

The purpose of this study was to assess the impact of Artificial Intelligence (AI) on education. Premised on a narrative and framework for assessing AI identified from a preliminary analysis, the scope of the study was limited to the application and effects of AI in administration, instruction, and learning. A qualitative research approach, leveraging the use of literature review as a research design and approach was used and effectively facilitated the realization of the study purpose. Artificial intelligence is a field of study and the resulting innovations and developments that have culminated in computers, machines, and other artifacts having human-like intelligence characterized by cognitive abilities, learning, adaptability, and decision-making capabilities. The study ascertained that AI has extensively been adopted and used in education, particularly by education institutions, in different forms. AI initially took the form of computer and computer related technologies, transitioning to web-based and online intelligent education systems, and ultimately with the use of embedded computer systems, together with other technologies, the use of humanoid robots and web-based chatbots to perform instructors' duties and functions independently or with instructors. Using these platforms, instructors have been able to perform different administrative functions, such as reviewing and grading students' assignments more effectively and efficiently, and achieve higher quality in their teaching activities. On the other hand, because the systems leverage machine learning and adaptability, curriculum and content has been customized and personalized in line with students' needs, which has

Source (pdf): IEEE Xplore Full-Text PDF:

How artificial intelligence will impact K-12 teachers

Existing and emerging technologies can help save teacher time—time that could be redirected toward student learning. But to capture the potential, stakeholders need to address four imperatives.

Source: Artificial intelligence in education: How will it impact K-12 teachers | McKinsey

An overview of some available Fairness Frameworks & Packages

Content

1. The LinkedIn Fairness Toolkit (LiFT)
2. Fairlearn: Fairness in machine learning mitigation algorithms
3. AI Fairness 360
4. Algofairness
5. FairSight: Visual Analytics for Fairness in Decision Making
6. GD-IQ: Spellcheck for Bias (code not available)
7. Aequitas: Bias and Fairness Audit Toolkit
8. 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
10. scikit-fairness
11. Mitigating Gender Bias In Captioning System

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**.

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.
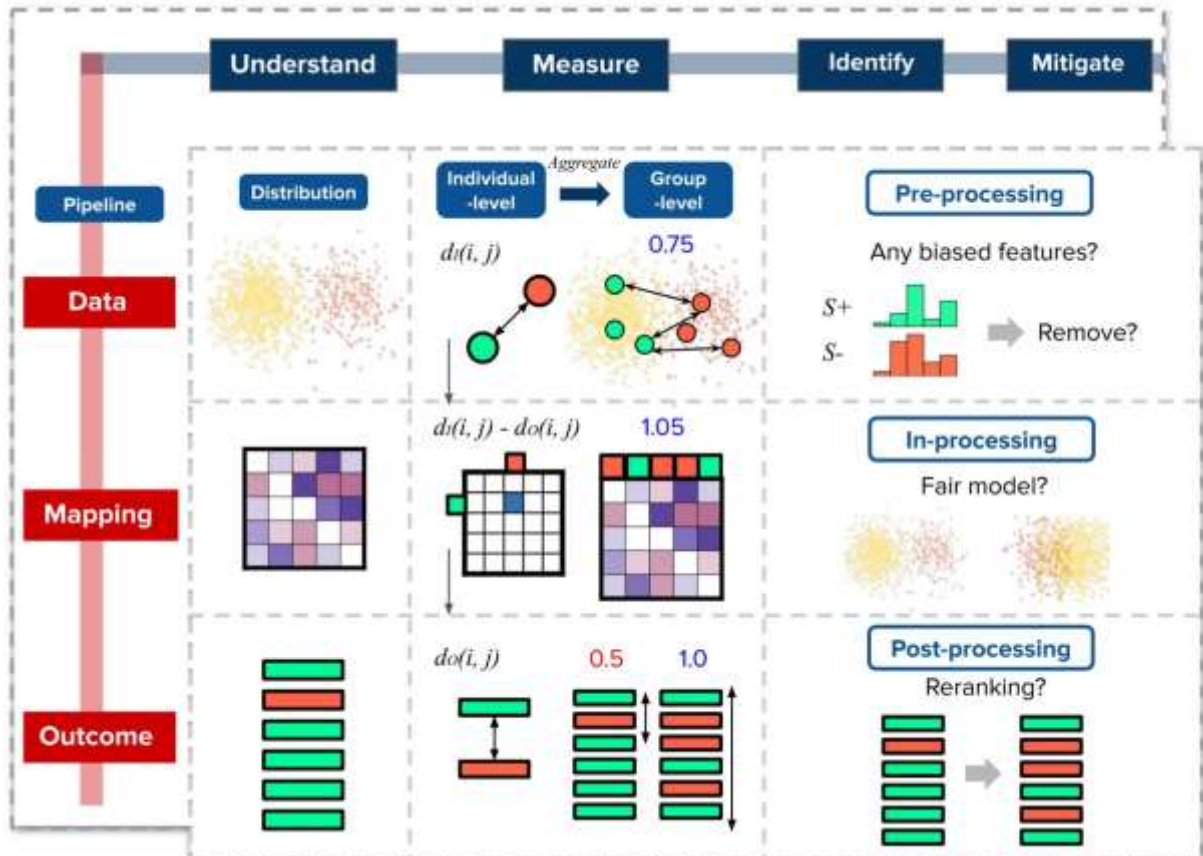
Specification

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.
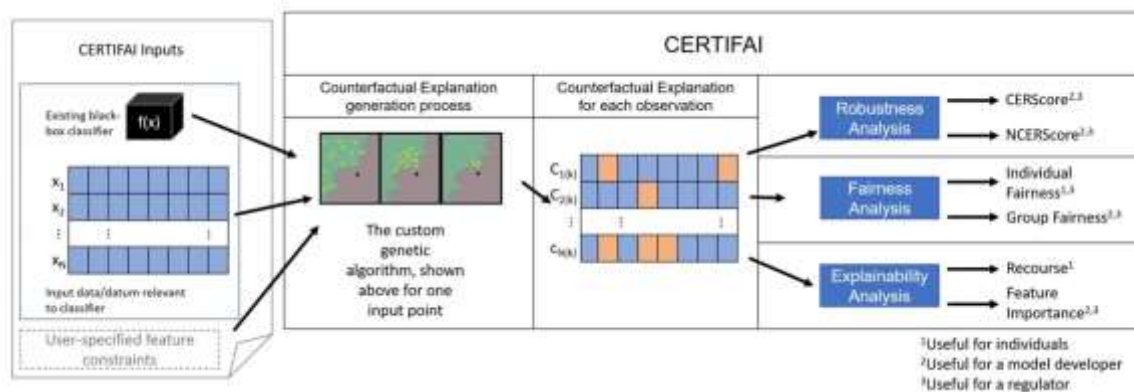

Figure 2: The CERTIFAI framework. Given a black-box ML model and input data along with optional user-specified feature constraints (such as feature type, range, etc.), the method generates counterfactual explanations using a genetic algorithm. The explanations can then be used for three purposes: explainability, 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.

DATA → PREPROCESS → MODEL → POST PROCESS → MEASURE

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.
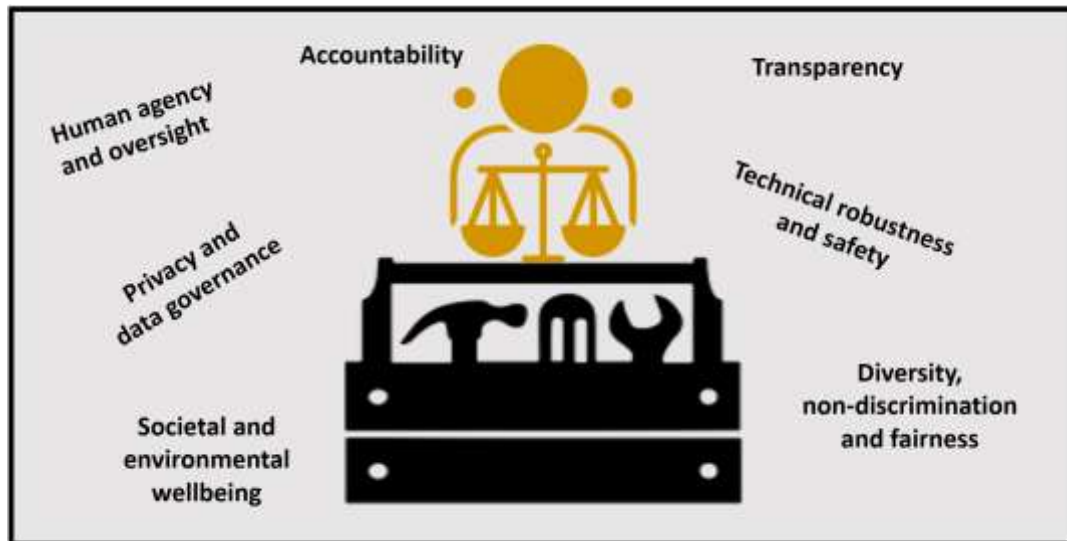
45
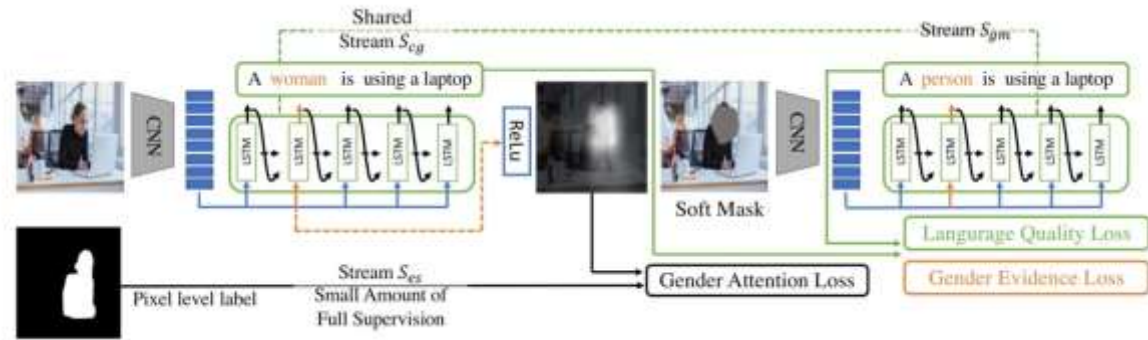
More info:

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 XAI**t

A brief overview of some
Ethical-AI Toolkits

audit-AI (Bias Testing for Generalized Machine Learning Applications)

audit-AI is a tool to measure and mitigate the effects of discriminatory patterns in training data and the predictions made by machine learning algorithms trained for the purposes of socially sensitive decision processes.

The overall goal of this research is to come up with a reasonable way to think about how to make machine learning algorithms more fair. While identifying potential bias in training datasets and by consequence the machine learning algorithms trained on them is not sufficient to solve the problem of discrimination, in a world where more and more decisions are being automated by Artificial Intelligence, our ability to understand and identify the degree to which an algorithm is fair or biased is a step in the right direction.

Link: https://github.com/pymetrics/audit-ai

Playing with AI Fairness (Google's new machine learning diagnostic tool lets users try on five different types of fairness)

Researchers and designers at Google's PAIR (People and AI Research) initiative created the What-If visualization tool as a pragmatic resource for developers of machine learning systems. Using the What-If tool reveals, however, one of the hardest, most complex, and most utterly human, questions raised by artificial intelligence systems: What do users want to count as fair?

link: https://pair-code.github.io/what-if-tool/ai-fairness.html

Ethics & Algorithms Toolkit (A risk management framework for governments)

Government leaders and staff who leverage algorithms are facing increasing pressure from the public, the media, and academic institutions to be more transparent and accountable about their use. Every day, stories come out describing the unintended or undesirable consequences of algorithms. Governments have not had the tools they need to understand and manage this new class of risk.

link: https://ethicstoolkit.ai/

AI Explainability 360 (IBM)

This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. We invite you to use it and improve it.

link: http://aix360.mybluemix.net/

PwC's Responsible AI

PwC's Responsible AI Toolkit is a suite of customizable frameworks, tools and processes designed to help you harness the power of AI in an ethical and responsible manner - from strategy through execution.

link: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html

InterpretML (Microsoft)

InterpretML is an open-source package that incorporates state-of-the-art machine learning interpretability techniques under one roof. With this package, you can train interpretable glassbox models and explain blackbox systems. InterpretML helps you understand your model's global behavior, or understand the reasons behind individual predictions.

Interpretability is essential for:

- Model debugging - Why did my model make this mistake?
- Detecting fairness issues - Does my model discriminate?
- Human-AI cooperation - How can I understand and trust the model's decisions?
- Regulatory compliance - Does my model satisfy legal requirements?
- High-risk applications - Healthcare, finance, judicial

link: https://github.com/interpretml/interpret

Diversity Toolkit: A Guide to Discussing Identity, Power and Privilege (USC)

This toolkit is meant for anyone who feels there is a lack of productive discourse around issues of diversity and the role of identity in social relationships, both on a micro (individual) and macro (communal) level. Perhaps you are a teacher, youth group facilitator, student affairs personnel or manage a team that works with an underserved population. Training of this kind can provide historical context about the politics of identity and the dynamics of power and privilege or help build greater self-awareness.

link: https://msw.usc.edu/mswusc-blog/diversity-workshop-guide-to-discussing-identity-power-and-privilege/#intro

WEF Empowering AI Leadership - An Oversight Toolkit for Boards of Directors (World Economic Forum)

This resource for boards of directors consists of: an introduction; 13 modules intended to align with traditional board committees, working groups and oversight concerns; and a glossary of artificial intelligence (AI) terms.

Eight modules focus on strategy oversight and the responsibilities connected with them. They cover: brand, competition, customers, operating model, people and culture, technology, cybersecurity and sustainable development. Five other modules cover additional board oversight topics: ethics, governance, risk, audit and board responsibilities.

Follow the links in the descriptions below to select a module. Continue scrolling down to read the introduction, which includes a description of AI.

Each module provides: a description of the topic, the board responsibilities specific to that module topic, the oversight tools, suggestions for setting an agenda for board discussions and resources for learning more about the topic.

link: https://adobe.ly/2WTBOmI

Responsible Innovation: A Best Practices Toolkit (Microsoft)

This toolkit provides developers with a set of practices in development, for anticipating and addressing the potential negative impacts of technology on people. We are sharing this as an early stage practice for feedback and learning.

Link: https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/

Bias and Fairness Audit Toolkit (Aequitas - Center for Data Science and Public Policy - University of Chicago)

The Bias Report is powered by Aequitas, an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.

Link: http://aequitas.dssg.io/

An Ethical Toolkit for Engineering/Design Practice (Markkula Center for Applied Ethics - Santa Clara University)

The tools below represent concrete ways of implementing ethical reflection, deliberation, and judgment into tech industry engineering and design workflows.

Used correctly, they will help to develop ethical engineering/design practices that are:

- Well integrated into the professional tech setting, and seen as a natural part of the job of good engineering and design (not external to it or superfluous)
- Made explicit so that ethical practice is not an 'unspoken' norm that can be overlooked or forgotten

- Regularized so that with repetition and habit, engineers/designers/technologists can gradually strengthen their skills of ethical analysis and judgment
- Operationalized so that engineers/designers are given clear guidance on what ethical practice looks like in their work setting, rather than being forced to fall back on their own personal and divergent interpretations of ethics

Link: https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

Fairlearn (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.

Link: https://github.com/fairlearn/fairlearn

TOOLBOX: Dynamics of AI Principles (AI ETHICS LAB)

You can use it to think through the ethical implications of the technologies

that you are evaluating or creating.

The Box is a simplified tool that

- lists important ethical principles and concerns,
- puts instrumental ethical principles in relation to core principles,
- helps visualize ethical strengths & weaknesses of technologies, and
- enables visual comparison of technologies.

Link: https://aiethicslab.com/big-picture/

Algorithmic Accountability Policy Toolkit (AI Now Institute)

The following toolkit is intended to provide legal and policy advocates with a basic understanding of government use of algorithms including, a breakdown of

key concepts and questions that may come up when engaging with this issue, an overview of existing research, and summaries of algorithmic systems currently used in government. This toolkit also includes resources for advocates interested in or currently engaged in work to uncover where algorithms are being used and to create transparency and accountability mechanisms

Link: https://ainowinstitute.org/aap-toolkit.pdf

From Principles to Practice – An interdisciplinary framework to operationalize AI ethics (AIEI Group)

The AI Ethics Impact Group is an interdisciplinary consortium led by VDE Association for Electrical, Electronic & Information Technologies and Bertelsmann Stiftung. We have come together in 2019 to bring AI ethics from principles to practice.

With our labelling and specification frameworks we aim to support the enforcement of European values and the protection of citizens in Europe, to create quality transparency and comparability in the market, and to prevent unnecessary red tape for companies with a straightforward implementation only where necessary.

We also want to make sure that AI ethics becomes easy to communicate and understand for organizations and citizens all over Europe and beyond.

We present our findings here and are looking forward to a broad implementation-oriented debate.

Link: https://www.ai-ethics-impact.org/en

Ethical OS Toolkit

What's in the Toolkit:

- A checklist of 8 risk zones to help you identify the emerging areas of risk and social harm most critical for your team to start considering now.
- 14 scenarios to spark conversation and stretch your imagination about the long-term impacts of tech you're building today.
- 7 future-proofing strategies to help you take ethical action today.

Link: https://ethicalos.org/

AI Philosophy

The Philosophy of AI and the AI of Philosophy (Professor John McCarthy)

Artificial intelligence (AI) has closer scientific connections with philosophy than do other sciences, because AI shares many concepts with philosophy, e.g. action, consciousness, epistemology (what it is sensible to say about the world), and even free will. This article treats the philosophy of AI but also analyzes some concepts common to philosophy and AI from the standpoint of AI.

Link: http://jmc.stanford.edu/articles/aiphil2.html

Philosophy will be the key that unlocks artificial intelligence (David Deutsch)

AI is achievable, but it will take more than computer science and neuroscience to develop machines that think like people

Link (The Guardian):
https://www.theguardian.com/science/2012/oct/03/philosophy-artificial-intelligence

Artificial Intelligence and Its Debt to the Mathematician Gödel (Javier Muñoz de la Cuesta)

It may be a good time to stop and think about it. In the end, developments developed by developments have elements that cannot be both consistent and complete without human involvement. It's down to us.

Link (Open Mind BBVA): Artificial Intelligence and Its Debt to the Mathematician Gödel (bbvaopenmind.com)

Aristotle and the chatbot: how ancient rules of logic could make artificial intelligence more human (David Ireland and Dana Bradford)

Many attempts to develop artificial intelligence are powered by powerful systems of mathematical logic. They tend to produce results that make logical sense to a computer program — but the result is not very human.
In our work building therapy chatbots, we have found using a different kind of logic — one first formalised by the Greek philosopher Aristotle more than 2,000 years ago — can produce results that are more fallible, but also much more like real people.

Link (The Conversation): Aristotle and the chatbot: how ancient rules of logic could make artificial intelligence more human (theconversation.com)

THE AI THOUGHT BOOK: Inspirational Thoughts & Quotes on Artificial Intelligence (Murat Durmus)

A somewhat different book about AI - GET INSPIRED!

"Quotations are drops of thoughts that sometimes reach deep into our soul and give us insights that a hundred books cannot." ~ Murat Durmus

Contents:

◆ Preface

◆ Artificial Intelligence

◆ AI-Ethics

◆ Explainable AI (XAI)

◆ Philosophy

◆ Data & Business

◆ Education & Future of Work

◆ Society & Humanity

◆ Mixed

◆ Three essays for the fundamental understanding of AI

◆ Epilog

◆ Appendix 1: Glossary

◆ Appendix 2: A brief History of Artificial Intelligence

◆ Appendix 3: The criminal Potential of AI

◆ Appendix 4: Some significant achievements in the field of AI since 2010

Link (ebook): Amazon.com: THE AI THOUGHT BOOK: Inspirational Thoughts & Quotes on Artificial Intelligence (including 13 colored illustrations & 3 essays for the fundamental understanding of AI) eBook: Durmus, Murat: Kindle Store

aisoma.de

Contact

Murat Durmus [in]

CEO & Founder @ AISOMA AG

Frankfurt am Main, Hessen, Deutschland · ·

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

murat.durmus@aisoma.de

https://www.aisoma.de