Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization

TAILOR
Grant Agreement Number 952215

D3.3 Handbook on Trustworthy AI Report (v1)

Document type (nature)	Report
Deliverable No	3.3
Work package number(s)	3
Date	Due M22, 30 June 2022
Responsible Beneficiary	CNR, ID 2
Author(s)	Umberto Straccia (CNR), Francesca Pratesi (CNR) For contributors, see the related section.
Publicity level	Public
Short description (Please insert the text in the Description of Deliverables in the Appendix 1.)	Handbook on Trustworthy AI

History			
Revision	**Date**	**Modification**	**Author**
1.0	2022-06-30	Version 1	Umberto Straccia

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**

Ana Paiva	8 / IST	30/06/2022
Koen van der Blom	7 / LEU	30/06/2022

The review documents are saved in the dedicated EMDESK folder.

Table of Contents

Summary of the report

WP3 decided to write the deliverable encyclopedia-like and present it in the form of a publically accessible WIKI. To do so, the Jupiter Book framework has been used

This report describes the initial structure and design of the TAILOR Handbook on Trustworthy AI wiki, which is currently and temporarily available at

https://prafra.github.io/jupyter-book-TAILOR-D3.2/TAILOR.html

An automatically generated pdf of the wiki is appended to this document.

The plan will be to integrate it into the TAILOR web page and to make a Wikipedia entry (by v2 of the handbook). A final paper book is also planned by then.

Introduction

The Handbook on Trustworthy AI assumes an encyclopedia-like structure and is presented in the form of a publically accessible WIKI. To do so, the Jupiter Book framework has been used.

In the long term, the handbook is meant to become a point of reference for resources (key concepts, tools, documentation, tutorials, teaching material, etc.) related to Trustworthy AI. The plan is also to integrate it into the TAILOR web page, and also to make a Wikipedia entry (by v2 of the handbook).

Each task leader of WP3 contributed to it and will update the content he is responsible for, as soon as major changes occur during the life period of the project.

An automatically generated pdf of the wiki is appended to this document.

This report describes the initial structure and design of the TAILOR Handbook on Trustworthy AI wiki, which is currently available at

https://prafra.github.io/jupyter-book-TAILOR-D3.2/TAILOR.html.

The link is temporary until the handbook will be included within the TAILOR web site.

Contributors

The following people have been involved in the Handbook on Trustworthy AI:

Partner ID / Acronym	**Name**	**Role**
2 / CNR	Umberto Straccia	Coordination, contributor
2 / CNR	Francesca Pratesi	Responsible for the whole Handbook wiki, responsible and contributor to the dimension “Explainable AI Systems”, contributor to the dimensions “Accountability and Reproducibility” and “Respect for Privacy” related respectively to Tasks 3.1, 3.4, and 3.5
2 / CNR	Riccardo Albertoni	Contributor to the dimensions “Accountability and Reproducibility” related to Task 3.4
2 / CNR	Sara Colantonio	Contributor to the dimensions “Accountability and Reproducibility” related to Task 3.4
1 / LiU	Fredrik Heintz	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
1 / LiU	Resmi Ramachandran Pillai	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
3 / INRIA	Guilherme Alves	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
3 / INRIA	Miguel Couceiro	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3

3 / INRIA	Karima Makhlouf	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
3 / INRIA	Sami Zhioua	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
4 / UCC	Gabriel Gonzalez-Castañé	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
4 / UCC	Andrea Rossi	Contributor to the dimension “Suistainability” related to Task 3.6
4 / UCC	Barry O’Sullivan	Contributor to the dimension “Suistainability” related to Task 3.6
4 / UCC	Andrea Visentin	Responsible and contributor to the dimension “Suistainability” related to Task 3.6
25 / TUD	Stefan Buijsman	Contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
25 / TUD	Sietze Kuilman	Contributor to the dimensions “Accountability and Reproducibility” related to Task 3.4
25 / TUD	Luciano C Siebert	Contributor to the dimensions “Accountability and Reproducibility” related to Task 3.4
25 / TUD	Arkady Zgonnikov	Contributor to the dimensions “Accountability and Reproducibility” related to Task 3.4
35 / PUT	Piotr Skrzypczyński	Contributor to the dimensions “Accountability and Reproducibility” related to Task 3.4
35 / PUT	Jerzy Stefanowski	Contributor to the dimensions “Accountability and Reproducibility” related to Task 3.4
40 / UNIPI	Riccardo Guidotti	Contributor to the dimension “Explainable AI Systems” related to Task 3.1
40 / UNIPI	Anna Monreale	Contributor to the dimension “Respect for Privacy” related to Task 3.5

40 / UNIPI	Roberto Pellungrini	Contributor to the dimension “Respect for Privacy” related to Task 3.5
40 / UNIPI	Salvatore Ruggieri	Responsible and contributor to the dimensions “Fairness, Equity, and Justice by Design” related to Task 3.3
41 / UGA	Marie-Christine Rousset	Responsible and contributor to the dimension “Respect for Privacy” related to Task 3.5
43 / UPV	Pablo A M Casares	Contributor to the dimension “Safety and Robustness” related to Task 3.2
43 / UPV	Santiago Escobar	Contributor to the dimension “Safety and Robustness” related to Task 3.2
43 / UPV	Jose Hernandez-Orallo	Responsible and contributor to the dimension “Safety and Robustness” related to Task 3.2
43 / UPV	Fernando Martinez-Plumed	Contributor to the dimension “Safety and Robustness” related to Task 3.2

Hosting

Currently and temporarily, the TAILOR Handbook on Trustworthy AI (v1), is hosted at

https://prafra.github.io/jupyter-book-TAILOR-D3.2/TAILOR.html

but will be integrated in the TAILOR web site (on-going).

Structure of the TAILOR Handbook on Trustworthy AI (v1)

Overall, the handbook content’s structure has been inspired by similar encyclopedia-like works such as the

Encyclopedia of Machine Learning and Data Mining. Editors: Claude Sammut, Geoffrey I. Webb, 2017. Springer. https://doi.org/10.1007/978-1-4899-7687-1

Besides an introductory part to TAILOR, for each dimension of Trustworthy AI,

- Explainable AI Systems
- Safety and Robustness
- Fairness, Equity, and Justice by Design

- Accountability and Reproducibility by design
- Respect for Privacy
- Sustainability

there are the following sections:

- Brief summary
- Abstract
- Motivation & Background
- Guidelines
- Software Frameworks Supporting Dimension
- Main Keywords
- List of authors that contributed to that dimension
- Bibliography

Each dimension has a series of entries associated. At the moment (i.e., in this first version) the number of entries ranges from 8 to 10 for each dimension.

Each entry has:

- Potential synonyms
- Brief summary
- A more detailed section
- Bibliography
- List of authors

There is also an

- Index

that lists all entries in alphabetical order, references to a short definition of an entry and where it is used within the handbook. Potential synonyms have their own entries in this index.

Appendix: PDF of the Handbook on Trustworthy AI (v1)

An automatically generated pdf of the handbook's wiki is included in here for completeness. However, the published version is the updated version.

Contenuti

Welcome to TAILOR

TAILOR: Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization

D3.3 Handbook on Trustworthy AI (Version 1)

Read this first

This is a working document for the Version 1 of the *D3.3 Handbook on Trustworthy AI*, the Tailor WP3 Handbook on Trustworthy AI. This is a Tailor project deliverable with two versions: Version 1 (M22) and Version 2 (M46).

About TAILOR

TAILOR is an EU-funded ICT-48 Network (GA 952215) with the purpose of building the capacity of providing the scientific foundations for Trustworthy AI in Europe by developing a network of research excellence centres leveraging and combining learning, optimization and reasoning.

- TAILOR will create a network of research excellence centres across all of Europe on the Foundations of Trustworthy AI based on four powerful instruments (a strategic roadmap committee, basic research program to address grand challenges, a connectivity fund for active dissemination to the larger AI community, and network collaboration activities promoting research exchanges, training materials and events, and joint PhD supervision.
- TAILOR will develop an ambitious research and innovation roadmap for the foundation of Trustworthy AI leveraging Europe's strengths and opportunities, across multiple disciplines, maturity levels, and geographical location. Seeds for its implementation will be proposed: challenges regarding both the basic research themes and application use-cases; a PhD program favouring immersion of PhDs in industry.
- TAILOR will launch and execute five basic research programs validating the operation of the network and performing ground-breaking basic research integrating learning, optimisation and reasoning in key areas for providing the scientific foundations for Trustworthy AI.
- TAILOR will develop and build on new mechanisms to step up AI outreach, harmonize training curricula, and significantly strengthen European capacities in AI research on Trustworthy AI.
- TAILOR brings together leading AI research centres from learning, optimisation and reasoning together with major European companies representing important industry sectors into a single scientific network to reduce the fragmentation, boost the collaboration, and increase the AI research capacity of Europe as well as attracting and retaining talents in Europe.

TAILOR, like all the research projects, is based on Work Pakages (WPs). WP3 (Trustworthy AI) aims at advancing knowledge on the six dimensions and putting each of them in relationships with foundation themes.

About the Encyclopedia

This book (to be consolidated in the second phase of the project) represents the first period deliverable, providing an encyclopedia of the major terms related to trustworthiness. Here, you can find definitions related to:

- Explainable AI.
- Safety and Robustness.
- Fairness, Equity, and Justice by Design.
- Accountability and Reproducibility.
- Respect for Privacy.
- Sustainability.

Complete List of Contributors

Coordinators:

- Umberto Straccia - Institute of Information Science and Technologies “A. Faedo” of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy
- Francesca Pratesi - Institute of Information Science and Technologies “A. Faedo” of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy

The complete list of authors is (in alphabetical order):

- Riccardo Albertoni - Istituto di Matematica Applicata e Tecnologie Informatiche “Enrico Magenes”, Consiglio Nazionale delle Ricerche (IMATI-CNR), Via De Marini, 6, 16149 Genova, Italy
- Tristan Allard - University of Rennes, CNRS, IRISA, 35000 Rennes, France
- Guilherme Alves, University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
- Alejandra Bringas Colmenarejo, School of Law, University of Southampton, SO17 1BJ, United Kindom
- Stefan Buijsman - Delft University of Technology, Jaffalaan 5, 2628 BX, Delft, The Netherlands
- Pablo A M Casares - VRAIN, Universitat Politècnica de València
- Sara Colantonio - Institute of Information Science and Technologies “A. Faedo” of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy
- Miguel Couceiro - Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
- Santiago Escobar - VRAIN, Universitat Politècnica de València
- Gabriel Gonzalez-Castañé - University College Cork, Cork, Ireland
- Riccardo Guidotti - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Fredrik Heintz - Linköping University, Department of Computer and Information Sciences, 58 183 Linköping, Sweden
- Jose Hernandez-Orallo - VRAIN, Universitat Politècnica de València
- Sietze Kuilman - Faculty of Electrical Engineering Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands
- Karima Makhlouf, Inria, Ecole Polytechnique, IPP, 91120, Paris, France
- Fernando Martinez-Plumed - VRAIN, Universitat Politècnica de València
- Anna Monreale - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Roberto Pellungrini - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Francesca Pratesi - Institute of Information Science and Technologies “A. Faedo” of the National Research Council of Italy (ISTI-CNR), Via G. Moruzzi, 1, 56124 Pisa, Italy
- Resmi Ramachandran Pillai - Linköping University, Department of Computer and Information Sciences, 58 183 Linköping, Sweden
- Andrea Rossi - SFI Centre for Research Training in Artificial Intelligence, University College Cork
- Marie-Christine Rousset - University of Grenoble Alpes, Grenoble, France
- Salvatore Ruggieri - University of Pisa, Department of Computer Sciences, Largo B. Pontecorvo, 3, 56127 Pisa, Italy
- Luciano C Siebert - Faculty of Electrical Engineering Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands
- Piotr Skrzypczyński - Institute of Robotics and Machine Intelligence, Poznań University of Technology, ul. Piotrowo 3A, 60-965 Poznań, Poland
- Jerzy Stefanowski - Institute of Computing Science, Poznań University of Technology, ul. Piotrowo 2, 60-965 Poznań, Poland
- Barry O’Sullivan - School of Computer Science & IT, University College Cork, Cork, Ireland
- Andrea Visentin - School of Computer Science & IT, University College Cork, Cork, Ireland
- Arkady Zgonnikov - Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, The Netherlands
- Sami Zhioua, Inria, Ecole Polytechnique, IPP, 91120, Paris, France

Explainable AI

In Brief

Explainable AI (often shortened to **XAI**) is one of the ethical dimensions that is studied in the TAILOR project. The origin of XAI dates back to the entering into force of the General Data Protection Regulation (GDPR). The GDPR [1], in its Recital 71, also mentions the right to explanation, as a suitable safeguard to ensure fair and transparent processing in respect of data subjects. It is defined as the right “to obtain an explanation of the decision reached after profiling”. According to NIST report [2], an explanation is the evidence, support, or reasoning related to a system’s output or process, where the output of a system differs by task, and the process refers to the procedures, design, and system workflow which underlie the system.

Abstract

While other aspects of ethics and trustworthiness, such as Respect for Privacy, are not novel concepts, and a lot of scientific literature has been explored on these topics, the study of explainability is a new challenge. In this part, we will cover the main elements that define the explanation of AI systems. First, we will try to survey briefly the main guidelines related to explainability. Then, we summarize a taxonomy that can be used to classify explanations. We will define the possible Dimensions of Explanations (e.g., we can discriminate between Model-Specific vs Model-Agnostic Explainers). Next, we will describe the requirements to provide good explanations and some of the problems related to the Explainability topic. Finally, we will give some examples of possible solutions we can adopt to provide explanations describing the reasoning behind an ML/AI model.

Motivation and Background

So far, the usage of black boxes in AI and ML processes has implied the possibility of inadvertently making wrong decisions due to a systematic bias in training data collection. Several practical examples have been provided, highlighting the “bias in, bias out” concept. One of the most famous examples of this concept regards a classification task: the algorithm’s goal was to distinguish between photos of Wolves and Eskimo Dogs (huskies) [1]. Here, the training phase of the process was done with 20 images, hand-selected such that all pictures of wolves had snow in the background, while pictures of huskies did not. This choice was intentional because it was part of a social experiment. In any case, on a collection of additional 60 images, the classifier predicts “Wolf” if there is snow (or light background at the bottom), and “Husky” otherwise, regardless of animal color, position, pose, etc (see an example in Fig. 1).

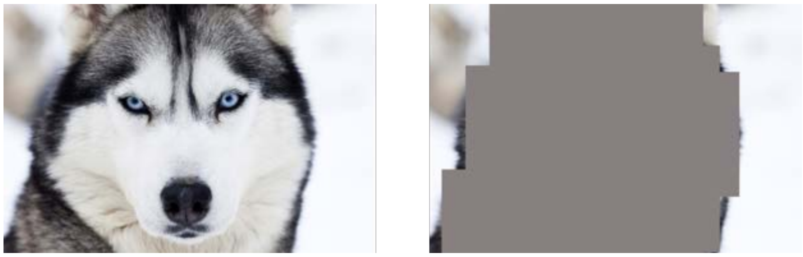


Fig. 1 Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task [1]. On the right, there is the original image; on the left, there is the explanation of the classification.

However, one of the most worrisome cases was discovered and published by ProPublica, an independent, nonprofit newsroom that produces investigative journalism with moral force. In [4], the authors showed how software can actually be racist. In a nutshell, the authors analyzed a tool called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions). COMPAS tries to predict, among other indexes, the recidivism of defendants, who are ranked low, medium, or high risk. It was used in many US states (such as New York and Wisconsin) to suggest to judges an appropriate probation or treatment plan for individuals being sentenced. Indeed, the tool was quite accurate (around 70 percent overall with 16,000 probationers), but ProPublica journalists found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.

From the above examples, it appears evident that explanation technologies can help companies for creating safer, more trustable products, and better managing any possible liability they may have.

Open the Black-Box Problem

The *Open the Black Box Problems* for understanding how a black box works can be summarized in the taxonomy proposed in [1] and reported in Fig. 2. The Open the Black Box Problems can be separated from one side as the problem of explaining how the decision system returned certain outcomes (*Black Box Explanation*) and on the other side as the problem of directly designing a transparent classifier that solves the same classification problem (*Transparent Box Design*). Moreover, the Black Box Explanation problem can be further divided among *Model Explanation* when the explanation involves the whole logic of the obscure classifier, *Outcome Explanation* when the target is to understand the reasons for the decisions on a given object, and *Model Inspection* when the target to understand how internally the black box behaves changing the input.

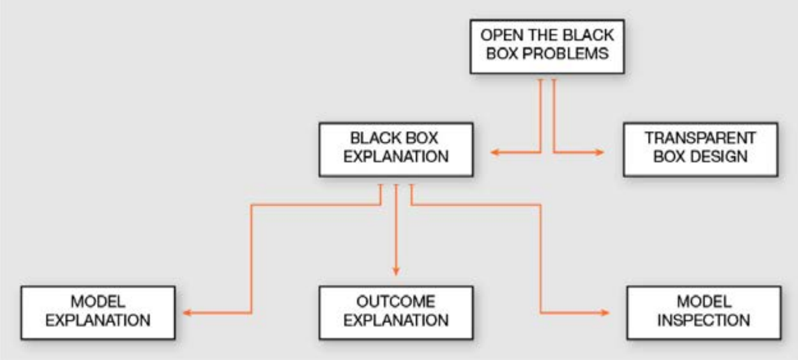


Fig. 2 A possible taxonomy about solutions to the Open the Black-Box problem [1].

On a different dimension, a lot of effort has been put into defining what are the possible Dimensions of Explanations (e.g., we can discriminate between Model-Specific vs Model-Agnostic Explainers), the requirements to provide good explanations (see guidelines), how to evaluate explanations and to understand the Feature Importance. Then, it is important to note that a variety of different kinds of explanations can be provided, such as Single Tree Approximation, feature_importance, Saliency Maps, Factual and Counterfactual, Exemplars and Counter-Exemplars, and Rules List and Rules Sets.

Guidelines

Given the relative novelty of the topic, a lot of guidelines have been developed in recent years.

However, the most authoritative guideline is the High-Level Expert Group on Artificial Intelligence - Ethics Guidelines for Trustworthy AI. Here, the explainability topic is included in the broader ./T3.1/transparency. According to this guideline, explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g., application areas of a system). Following the GDPR interpretation, in [1], it is stated that whenever an AI system has a significant impact on people’s lives, it should be possible to demand a suitable explanation of the AI system’s decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g., layperson, regulator, or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

Another distinguished authority that has been worked on ethical guidance is **the Alan Turing Institute**, the UK’s national institute for data science and artificial intelligence, where David Leslie [7] summarized the risks due to the lack of transparency or the absence of a valid explanation, and he advocates the use of ./T3.1/counterfactuals for contrasting unfair decisions. Together with the Information Commissioner’s Office (ICO), which is responsible for overseeing data protection in the UK, it has been published more recent and complete guidance [2]. Here, six steps are recommended to develop a system:

1. Select priority explanations by considering the domain, use case, and impact on the individual.
2. Collect and pre-process data in an explanation-aware manner, stressing the fact that the way in which data is collected and pre-processed may affect the quality of the explanation.
3. Build systems to ensure to being able to extract relevant information for a range of explanation types.
4. Translate the rationale of your system's results into useable and easily understandable reasons, e.g., transforming the model's logic from quantitative rationale into intuitive reasons or using everyday language that can be understood by non-technical stakeholders.
5. Prepare implementers to deploy the AI system, through appropriate training.
6. Consider how to build and present the explanation, particularly keeping in mind the context and contextual factors (domain, impact, data, urgency, audience) to deliver appropriate information to the individual.

Nevertheless, the attention on this theme is not relegated to the European border. Indeed, as an example of US effort in dealing with Explainability and Ethics, **NIST, the National Institute of Standards and Technology of Maryland**, developed some guidelines, and a white paper [2] was published after a first draft[1] was published in 2020, a variety of comments[2] was collected, and a workshop[3] involving different stakeholders was held. The white paper [2] analyzes the multidisciplinary nature of explainable AI and acknowledges the existence of different users who requires different kinds of explanations, stating that one-size-fits-all explanations do not exist. The fundamental properties of explanations contained in the report are:

- *Meaningfulness*, i.e., explanations must be understandable to the intended consumer(s). This means that there is the need to consider the intended audience and some characteristics they can have, such as prior knowledge or the overall psychological differences between people. Moreover, the explanation's purpose is relevant too. Indeed, different scenarios and needs impact on what is important and useful in a given context. This implies understanding the audience's needs, level of expertise, and relevancy to the question or query.
- *Accuracy*, i.e., explanations correctly reflect a system's process for generating its output. Explanation accuracy is a distinct concept from decision accuracy. Explanation accuracy needs to account for the level of detail in the explanation. This second principle might be in contrast with the previous one: a detailed explanation may accurately reflect the system's processing but sacrifice how useful and accessible it is to certain audiences, while a brief, simple explanation may be highly understandable but would not fully characterize the system.
- *Knowledge limits*, i.e., characterizing the fact that a system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output. This practice safeguard answers so that a judgment is not provided when it may be inappropriate to do so. This principle can increase trust in a system by preventing misleading, dangerous, or unjust outputs.

Software Frameworks Supporting Dimension

Within the European Research Council (ERC) XAI project and the European Union's Horizon 2020 SoBigData++ project, we are developing an infrastructure for sharing experimental datasets and explanation algorithms with the research community, creating a common ground for researchers working on explanations of black boxes from different domains. All resources, provided they are not prohibited by specific legal/ethical constraints, will be collected and described in a findable catalogue. A dedicated virtual research environment will be activated, so that a variety of relevant resources, such as data, methods, experimental workflows, platforms, and literature, will be managed through the SoBigData++ e-infrastructure services and made available to the research community through a variety of regulated access policies. We will provide a link to the libraries and framework as soon as they be will fully published.

Main Keywords

- Kinds of Explanations: Explanations returned by an AI system depend on various factors (such as the task or the available data); generally speaking, each kind of explanations serves better a specific context.
- Feature Importance: The **feature importance** technique provides a score, representing the "importance", for all the input features for a given AI model, i.e., a higher importance means that the corresponding feature will have a larger effect on the model.

- Saliency Maps: Saliency maps are explanations used on image classification tasks. A **saliency map** is an image where each pixel's color represents a value modeling the importance of that pixel in the original image (i.e., the one given in input to the explainer) for the prediction.
- Single Tree Approximation: The **single tree appoximation** is an approach that aims at building a decision tree to approximate the behavior of a black box, typically a neural network.
- Dimensions of Explanations: **Dimensions of explanations** are useful to analyze the interpretability of AI systems and to classify the explanation method.
- Black Box Explanation vs Explanation by Design: The difference between **Black Box Explanation** (or **Post-hoc Explanations**) and **Explanation by Design** (or **Ante-hoc Explanations**) regards the ability to know and exploit the behaviour of the AI model. With a black box explanation, we pair the black box model with an interpretation the black box decisions or model, while in the second case, the strategy is to rely, by design, on a transparent model.
- Model-Specific vs Model-Agnostic Explainers: We distinguish between **model-specific** or **model-agnostic** explanation method depending on whether the technique adopted to retrieve the explanation acts on a particular model adopted by an AI system, or can be used on any type of AI.
- Global vs Local Explanations: We distinguish between a **global** or **local** explanation depending on whether the explanation allows understanding the whole logic of a model used by an AI system or the explanation refers to a specific case, i.e., only a single decision is interpretable.

Bibliography

[[1]] European Parliament & Council. General Data Protection Regulation. 2016. L119, 4/5/2016, p. 1–88.

[2](1,2,3) P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, and Mark A. Przybocki. Four principles of explainable artificial intelligence. NISTIR 8312, September 2021. URL: https://doi.org/10.6028/NIST.IR.8312 (visited on 2022-02-16).

[3](1,2) M.T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": explaining the predictions of any classifier. In *SIGKDD*. 2016.

[[4]] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 2020-09-22).

[5](1,2) R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 2018.

[[6]] High Level Expert Group on AI. Ethics Guidelines for Trustworthy AI. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on 2022-05-10).

[[7]] David Leslie (the Alan Turing Institute). Understanding artificial intelligence ethics and safety - a guide for the responsible design and implementation of ai systems in the public sector. URL: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf (visited on 2022-02-16).

[[8]] Information Commissioner's Office (ICO). Accountability framework. URL: https://ico.org.uk/for-organisations/accountability-framework/ (visited on 2022-05-25).

This entry was readapted from *Pratesi, Trasarti, Giannotti. Ethics in Smart Information Systems. Policy Press (currently under review)* and from *Guidotti, Monreale, Ruggieri, Turini, Giannotti, Pedreschi. A survey of methods for explaining black box models. ACM Computing Surveys, Volume 51 Issue 5 (2019)* by Francesca Pratesi.

[[1]] https://doi.org/10.6028/NIST.IR.8312-draft

[[2]] https://www.nist.gov/artificial-intelligence/comments-received-four-principles-explainable-artificial-intelligence-nistir

[[3]] https://www.nist.gov/system/files/documents/2021/09/24/XAI_Workshop_Summary_Final_20210922.pdf

Kinds of Explanations

In brief

Explanations returned by an AI system depend on various factors (such as the task or the available data); generally speaking, each kind of explanations serves better a specific context.

More in detail

Increasing research on XAI is bringing to light a wide list of explanations and explanation methods for “opening” black box models. The explanations returned depend on various factors, such as:

- the type of task they are needed for,
- on which kind of data the AI system acts,
- who is the final user of the explanation,
- if they allow to explain the whole behavior of the AI system (global explanations) or reveal the reasons for the decision only for a particular instance (local explanations),
- the business perspective, i.e., which are the implication of companies in having explainable and interpretable systems and models, in terms of business strategies and secrecy,
- the fact that, in a decentralized node, an explanation could require information that is nor directly available on site.

In this part of the Encyclopedia, we review a subset of the most used types of explanations and show how some state-of-the-art explanation methods can return them. The interested reader can refer to [2], [1] for a complete review of XAI literature.

Bibliography

[[1]] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 2018.

[[2]] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

Feature Importance

In brief

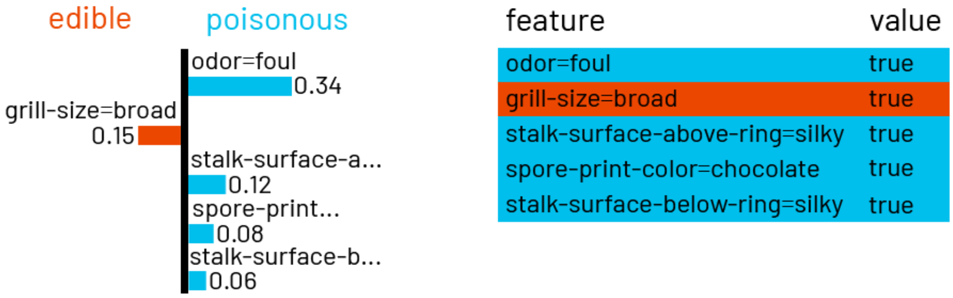
The **feature importance** technique provides a score, representing the “importance”, for all the input features for a given AI model, i.e., a higher importance means that the corresponding feature will have a larger effect on the model.

More in detail

Local explanations, especially when dealing with tabular data, can also be returned in the form of **features importance**. This technique considers both the sign and the magnitude of the contribution of the features for a given AI decision. In particular, if the value of a feature is positive, then it contributes by increasing the model’s output; at the same time, if the sign is negative, then the feature decreases the output of the model. At the same time, if a feature has a higher contribution than another, both positively and negatively, then it means that it has a stronger influence on the prediction of the black box outcome. The magnitude of the provided score expresses this meaning.

The features importance summarizes the outcome of the black box model, allowing quantifying the changes of the black box decision for each test record. This means it is possible to identify the features leading to a specific outcome for a certain instance and how much they contributed to the decision. In the following, we provide a couple of examples obtained with two of the most popular methods, namely LIME and SHAP, both model-agnostic local explanation methods.

The LIME explanation method [1] randomly generates synthetic instances around the analyzed record. Then it returns the features importance as the coefficient of a linear regression model adopted as a local surrogate. The synthetic instances are weighted according to their proximity to the instance of interest. The Lasso model is trained to approximate the probability of the decision of the black box in the synthetic neighborhood of the instance analyzed. Fig. 3 shows the features importance returned by LIME in a scenario where we want to predict if a mushroom is edible or poisonous. Indeed, here the prediction class has only two classes, and we want to understand which features are most relevant to discriminate between the two classes. In this example, the feature *odor=foul* has a positive contribution of 0.26 in predicting of a mushroom as *poisonous*, *stack-surface-above-ring=silky* has a positive contribution of 0.11, *spore-print-color=chocolate* has a positive contribution of 0.08, *stack-surface-below-ring=silky* has a positive contribution of 0.06, while *gill-size=broad* has a negative contribution of 0.13.



feature	value
odor=foul	true
grill-size=broad	true
stalk-surface-above-ring=silky	true
spore-print-color=chocolate	true
stalk-surface-below-ring=silky	true

Fig. 3 Example of explanation based on features importance by LIME [1].

The SHAP explanation method [3] connects game theory with local explanations exploiting the *Shapely values* of a conditional expectation function of the black box to explain the AI. Shapley values are introduced in [4] with a method for assigning “payouts” to “players” depending on their contribution to the “total payout”. Players cooperate in a coalition and receive a certain “profit” from this cooperation. The connection with explainability is as follows. The “game” is the decision of the black box for a specific instance, while the “profit” is the actual value of the decision for this instance minus the average values for all instances. The “players” are the feature values of the instance that leads towards a certain value, i.e., collaborate to receive the profit. Thus, a Shapley value is the *average marginal contribution* of a feature value across all possible coalitions, i.e., combinations [5]. Therefore, SHAP returns the local unique additive feature importance for each specific record. The higher a Shapely value, the higher the contribution of the feature. Fig. 4 shows an example of SHAP explanation, where the features importance is expressed in the form of a *force plot*. The example is based on the Boston Housing Dataset[1], a dataset that collects information relative to a portion of US census data (such as the age) along with some information about the areas (i.e., pupil-teacher ratio by town or accessibility to radial highways). This explanation each feature’s contribution levelin pushing the black box prediction from the base value (the average model output over the dataset, which is 24.41 in this example) to the model output. The features pushing the prediction higher are shown in light blue, and those pushing the prediction lower are shown in orange.

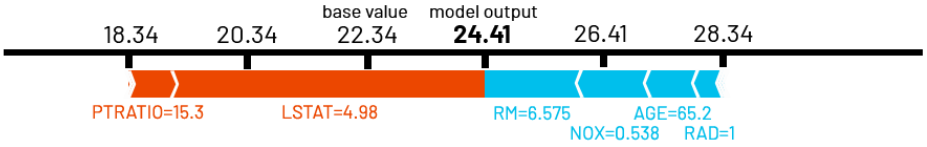


Fig. 4 Example of explanation based on features importance by SHAP [1].

Even if the classic application case of feature importance is with tabular data, it is important to note that, under appropriate settings, LIME and SHAP can also be used to explain the decisions of AI working on textual data and images.

Bibliography

[[1]] M.T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": explaining the predictions of any classifier. In *SIGKDD*. 2016.

[2](1,2) Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. *Principles of Explainable Artificial Intelligence*. Springer International Publishing, 2021.

[[3]] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774. 2017.

[[4]] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[[5]] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

[[1]] https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

Saliency Maps

In brief

Saliency maps are explanations used on image classification tasks. A **saliency map** is an image where each pixel's color represents a value modeling the importance of that pixel in the original image (i.e., the one given in input to the explainer) for the prediction.

More in detail

The most used type of explanation for explaining AI systems working on images consists of **saliency maps**. A saliency map is an image where each pixel's color represents a value modeling the importance of that pixel for the prediction, i.e., they show the positive (or negative) contribution of each pixel to the black box outcome. Saliency maps are a very typical example of local explanation methods since they are tailored to the image that must be explained.

In the literature, there exist different explanation methods locally explaining deep neural networks for image classification. The two most used model-specific techniques are *perturbation-based attribution methods* [3, 4] and *gradient attribution methods* such as SAL [5], ELRP [6], GRAD [7], and INTG [8].

Without entering into the details, these XAI approaches aim at attributing an importance score to each pixel in order to minimize the probability of the deep neural network (DNN) labeling the image with a different outcome when only the most important pixels are considered. Indeed, the areas retrieved by these methods are also called *attention areas*.

The aforementioned XAI methods are specifically designed for specific DNN models, i.e., they are model-specific.

However, relying on appropriate image transformations that take advantage of the concept of “superpixels” [1], i.e., the results of the segmentation of an image into regions by considering proximity or similarity measures, also model-agnostic explanation methods (such as LIME [1], ANCHOR [9], and LORE [10]) can be employed to explain AI working on images for any kind of black box model.

The attention areas of explanations returned by these methods are strictly dependent on both:

- the technique used for segmenting the image to explain and
- to a neighborhood consisting of unrealistic synthetic images with “suppressed” superpixels [11].

A different approach for generating neighborhoods is introduced by the local model-agostic explanation method ABELE [12]. This method relies on a generative model, i.e., an adversarial autoencoder [13], to produce a realistic synthetic neighborhood that allows retrieving more understandable saliency maps. Indeed, saliency maps returned by ABELE highlight the contiguous attention areas that can be varied while maintaining the same classification from the black box used by the AI system.

Fig. 5 reports a comparison of saliency maps for the classification of the handwritten digits “9” and “0” for the explanation methods ABELE [12, 14], LIME [1], SAL [5], ELRP [6], GRAD [7], and INTG [8].

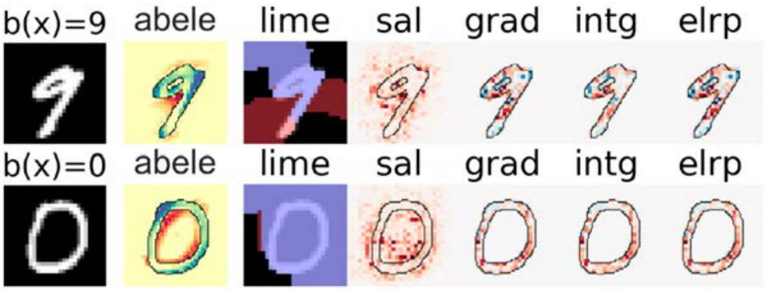


Fig. 5 Example of saliency maps returned by different explanation methods. The first column contains the image analyzed and the label assigned by the black box model *b* of the AI system. [1].

Bibliography

[1](1,2,3) M.T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": explaining the predictions of any classifier. In *SIGKDD*. 2016.

[[2]] Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. *Principles of Explainable Artificial Intelligence*. Springer International Publishing, 2021.

[[3]] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3429–3437. 2017.

[[4]] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer, 2014.

[5](1,2)

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[6](1,2) Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[7](1,2) Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[8](1,2) Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[[9]] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: high-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[[10]] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. 2018. https://arxiv.org/abs/1805.10820.

[[11]] Riccardo Guidotti, Anna Monreale, and Leonardo Cariaggi. Investigating neighborhood generation methods for explanations of obscure image classifiers. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 55–68. Springer, 2019.

[12](1,2) Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. Black box explanation by learning image exemplars in the latent feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 189–205. Springer, 2019.

[[13]] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[[14]] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

Single Tree Approximation

In brief

The **single tree appoximation** is an approach that aims at building a decision tree to approximate the behavior of a black box, typically a neural network.

More in detail

Decision trees are the simplest example of transparent techniques. Moreover, they can be built to provide post-hoc explanations of black-boxes. One of the first approaches introduced to explain neural networks is TREPAN [2]. TREPAN is a global explanation method that is able to model the whole logic of a neural network working on tabular data with a **single decision tree** [3]. The decision tree returned by TREPAN as an explanation is a *global transparent surrogate*. Indeed, every path from the root of the tree to a leaf explains the reasons for the final decision that is reported in the leaf itself. An example of a decision tree returned by TREPAN is illustrated in Fig. 6.

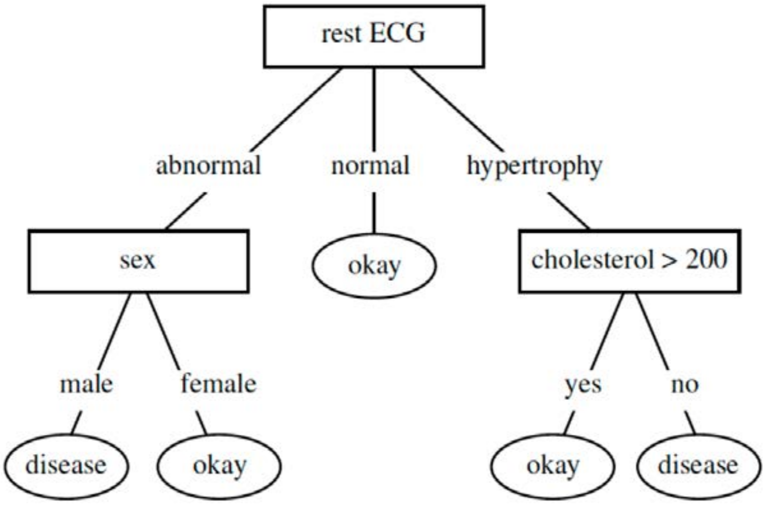


Fig. 6 Example of global tree-based explanation returned by TREPAN [1].

This global explanation reveals that the black box first focuses on the value of the feature *rest ECG*; Depending on its degree (abnormal, normal, hypertrophy), tha black box takes different decisions depending on additional factors such as sex or cholesterol. In particular, TREPAN queries the neural network to induce the decision tree by maximizing the *gain ratio* [3] on the data with respect to the neural network's predictions.

A weakness of common decision trees like ID3 or C4.5 [4] is that the amount of data to find the splits near to the leaves is much lower than those used initially. Thus, in order to retrieve how a neural network works in detail, TREPAN adopts a synthetic generation of data that respects the path of each node before performing the splitting such that the same amount of data is used for every split. In addition, it allows flexibility by using *"m-of-n" rules* where only *m* conditions out of *n* are required to be satisfied to classify a record. Therefore, TREPAN maximizes the fidelity of the single tree explanation with respect to the black box decision.

It is worth noting that, even though TREPAN is proposed to explain neural networks, in reality it is model-agnostic because it does not exploit any internal characteristic of neural networks to retrieve the explanation tree. Moreover, it does not place any requirements on either the architecture of the network or its training method. Thus, it can be theoretically employed to provide explanations to every kind of classifier.

In [5] is presented an extension of TREPAN that aims to keep the tree explanation simple and compact by introducing four splitting approaches in order to find the most relevant features during the tree construction. In [6], genetic programming is used for building a single decision tree that approximates the behavior of a neural network ensemble by considering additional genetic features obtained as combinations of the original data and the novel data annotated by the black box models. Both methods described in [5], [6], like TREPAN, return explanations in the form of a global decision tree.

Bibliography

[[1]] Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. *Principles of Explainable Artificial Intelligence*. Springer International Publishing, 2021.

[[2]] M. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, volume 8, 24–30. 1996.

[3](1,2) Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Addison-Wesley, 2006.

[[4]] J Ross Quinlan. *C4.5: Programs for Machine Learning*. Elsevier, 1993.

[5](1,2) Olcay Boz. Extracting decision trees from trained neural networks. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 456–461. 2002.

[6](1,2) Ulf Johansson and Lars Niklasson. Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 238–244. IEEE, 2009.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

Dimensions of Explanations

In brief

Dimensions of Explanations are useful to analyze the interpretability of AI systems and to classify the explanation method.

More in detail

The goal of Explainable AI is to *interpret* AI reasoning. According to Merriam-webster, to *interpret* means to give or provide the meaning or to explain and present in understandable terms some concepts. Therefore, in AI, *interpretability* is defined as the ability to *explain* or to provide the meaning in understandable terms to a human [4],[5]. These definitions assume that the concepts composing an explanation and expressed in understandable terms are self-contained and do not need further explanations. An explanation is an “interface” between a human and an AI, and it is simultaneously both human understandable and an accurate proxy of the AI.

We can identify a set of dimensions to analyze the interpretability of AI systems that, in turn, reflect on existing different types of explanations [1]. Some of these dimensions are related to *functional requirements* of explainable artificial intelligence, i.e., requirements that identify the algorithmic adequacy of a particular approach for a specific application, while others to the *operational requirements*, i.e., requirements that take into consideration how users interact with an explainable system and what is the expectation. Some dimensions instead derive from the need for *usability criteria* from a user perspective, while others derive from the need for guarantees against any vulnerability issues.

Recently, all these requirements have been analyzed [6] to provide a framework allowing the systematic comparison of explainability methods. In particular, in [6], Sokol and Flach propose *Explainability Fact Sheets*, which enable researchers and practitioners to assess capabilities and limitations of a particular explainable method. As an example, given an explanation method *m*, we can consider the following functional requirements.

- *(i)* Even though *m* is designed to explain regressors, can we use it to explain probabilistic classifiers?
- *(ii)* Can we employ *m* on categorical features even though it only works on numerical features? On the other hand, as an operational requirement, can we consider which is the *function of the explanation*? Provide transparency, assess the fairness, etc.

Besides the detailed requirements illustrated in [6], in the literature it is recognized a categorization of explanation methods among fundamental pillars [2],[1]:

- *(i)* Black Box Explanation vs Explanation by Design,
- *(ii)* Global vs Local Explanations,
- *(iii)* Model-Specific vs Model-Agnostic Explainers.

Fig. 7 illustrates a summarized ontology of the taxonomy used to classify XAI methods.

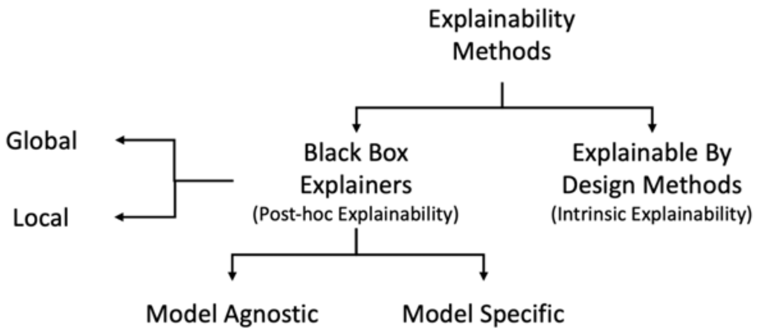


Fig. 7 A summarized ontology of the taxonomy of XAI methods [1].

Bibliography

[1](1,2) R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 2018.

[[2]] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[[3]] Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. *Principles of Explainable Artificial Intelligence*. Springer International Publishing, 2021.

[[4]] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, and R. Benjamins. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. In *Information Fusion*. 2020.

[[5]] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[6](1,2,3) Kacper Sokol and Peter A. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *ACM Conference on Fairness, Accountability, and Transparency*, 56–67. ACM, 2020.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

Black Box Explanation vs Explanation by Design

Synonyms: Post-hoc vs Ante-hoc Explanations.

In brief

The difference between **Black Box Explanation** (or **Post-hoc Explanations**) and **Explanation by Design** (or **Ante-hoc Explanations**) regards the ability to know and exploit the behaviour of the AI model. With a black box explanation, we pair the black box model with an interpretation the black box decisions or model, while in the second case, the strategy is to rely, by design, on a transparent model.

More in detail

When we talk about **Black Box Explanation**, the strategy is to couple an AI with a black box model with an explanation method able to interpret the black box decisions. In the case of **Explanation by Design** (aka **Transparency**), the idea is to substitute the obscure model with a transparent model in which the decision process is accessible by design, i.e., explainability is inserted into a model from the very beginning.

Figure 8 depicts this distinction.

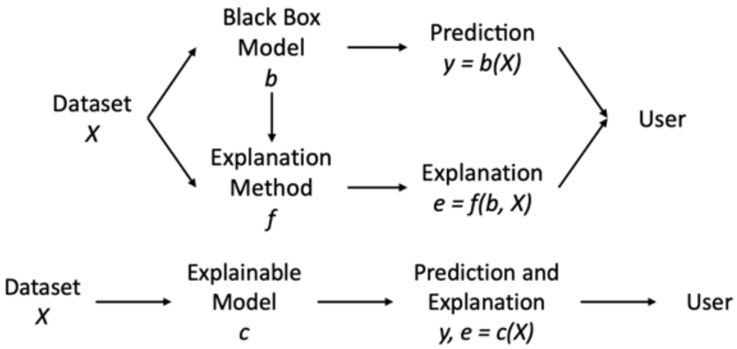


Fig. 8 (Top) Black box explanation pipeline. (Bottom) Explanation by design pipeline. [1].

Starting from a dataset *X*, the **black box explanation** idea is to maintain the high performance of the obscure model *b* used by the AI, which is allowed to be trained normally, and to use an explanation method *f* to retrieve an explanation *e* by reasoning over *b* and *X*. In such a way, we aim to reach both accuracy and the ability to gain some Kinds of Explanations. This kind of approach is the one more addressed nowadays in the XAI research field [1] [2] [3].

On the other hand, the **explanation by design** consists of directly designing a comprehensible model *c* over the dataset *X*, which is interpretable by design and returns an explanation *e* besides the prediction *y*. Thus, the idea is to use this transparent model directly in the AI system [5] [6]. In the literature, there are various models recognized to be interpretable. Examples include decision trees, decision rules, and linear models [7]. These models are considered easily understandable and interpretable for humans. However, nearly all of them sacrifice performance in favor of interpretability. In addition, they cannot be applied effectively on data types such as images or text, but only on tabular, relational data, i.e., tables.

Bibliography

[[1]] M.T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": explaining the predictions of any classifier. In *SIGKDD*. 2016.

[[2]] Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. *Principles of Explainable Artificial Intelligence*. Springer International Publishing, 2021.

[[3]] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774. 2017.

[[4]] M. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, volume 8, 24–30. 1996.

[[5]] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[[6]] C. Rudin and J. Radin. Why are we using black box models in ai when we don’t need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 2019.

[[7]] A. A. Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

Model-Specific vs Model-Agnostic Explainers

Synonyms: Not Generalizable vs Generalizable Explanations.

In brief

We distinguish between **model-specific** or **model-agnostic** explanation methods depending on whether the technique adopted to retrieve the explanation acts on a particular model adopted by an AI system, or can be used on any type of AI.

More in detail

This is one of the first Dimensions of Explanations we should consider.

The most used approach to explain AI black boxes is known as *reverse engineering*. The name comes from the fact that the explanation is retrieved by observing what happens to the output, i.e., the AI decision, when changing the input in a controlled way. An explanation method is model-specific, or not generalizable [2], whether it considers inputs or outputs as well as the inner-workings of a machine learning model. The drawback of this approach is that it can be used to interpret only particular types of black box models. For example, if an explanation approach is designed to interpret a Random Forest [3] and internally uses a concept of distance between trees, then such an approach cannot be used to explain the predictions of a neural network.

On the other hand, an explanation method is model-agnostic, or generalizable, when it can be used independently from the black box model being explained. In other words, the AI's internal characteristics are not exploited to build the interpretable model approximating the black box behavior.

In Fig. 9, there is a summarization of these two kinds of approaches.

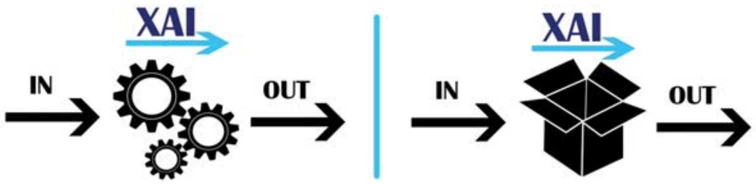


Fig. 9 On the left, an explainer which exploites the internal structure and behaviour of the model. On the right, an explainer which uses the model as a black-box to understand its reasoning.

Bibliography

[[1]] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Addison-Wesley, 2006.

[[2]] David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183((3)):1466—1476, 2007.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

Global vs Local Explanations

In brief

We distinguish between a **global** or **local** explanation depending on whether the explanation allows understanding the whole logic of a model used by an AI system or the explanation refers to a specific case, i.e., only a single decision is interpretable.

More in detail

A **global explanation** consists in providing a way for interpreting any possible decision of a *black box model*. Generally, the black box behavior is approximated with a transparent model trained to mimic the obscure model (see Black Box Explanation vs Explanation by Design) and also to be human-understandable. In other words, the interpretable model approximating the black box provides a global interpretation. Unfortunately, global explanations are quite hard to achieve and, up to now, can be provided only for AI working on relational data.

A **local explanation** consists in retrieving the reasons for a specific *outcome* returned by a black box model relatively to the decision for a certain instance. In this case, it is not required to explain the whole logic underlying the AI, but a local explanation only provides the reason for the prediction on a specific input instance. Hence, an interpretable model is used to approximate the black box behavior only in the "neighborhood" of the instance analyzed, i.e., with respect to similar instances only. The idea is that it is easier to approximate the AI with a simple and understandable model, in such a neighborhood. Regarding Figure 8 (top) (see Black Box Explanation vs Explanation by Design), a *global* explanation method *f* uses many instances *X* over which the explanation is returned.

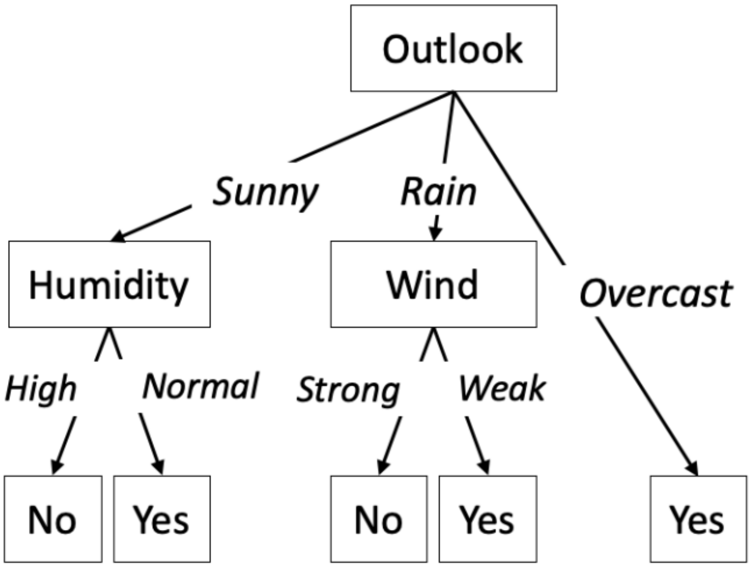
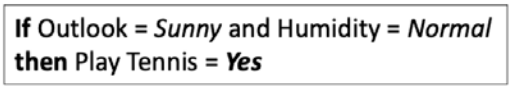


Fig. 10 Global explanation example in the form of decision tree [1].

Figure 10 illustrates an example of a global explanation *e* obtained by a decision tree structure for a classifier recommending to play tennis or not. The overall decision logic is captured by the tree that says that the classifier recommends playing tennis or not by first looking at the *Outlook* feature. If its values *Overcast*, then the prediction is "not to play". Otherwise, if its value is *Sunny*, the classifier checks the *Humidity* feature and recommends "not to play" if the *Humidity* is *High* and "to play" if it is *Normal*. The same reasoning applies to the other branch of the tree. Again, with reference to Figure 8 (top) (see Black Box Explanation vs Explanation by Design), a local explanation method *f* returns an explanation only for a single instance *x*.

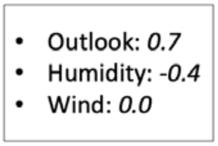
Two examples of local explanations are given in the following.



If Outlook = ***Sunny*** and Humidity = ***Normal***
then Play Tennis = ***Yes***

Fig. 11 Local explanation example in the form of decision rule [1].

The local rule-based explanation (Figure 11) *e* for a given record *x* says that the black box *b* suggested playing tennis because the *Outlook* is *Sunny* and the *Humidity* is *Normal*.



- Outlook: *0.7*
- Humidity: *-0.4*
- Wind: *0.0*

Fig. 12 Local explanation example in form of feature importance [1].

On the other hand, the explanation *e* formed by features importance (Figure 12) says that the black box *b* suggested playing tennis because the *Outlook* has a large positive contribution, *Humidity* has a consistent negative contribution, and *Wind* has no contribution in the decision.

Bibliography

[1](1,2,3) Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. *Principles of Explainable Artificial Intelligence*. Springer International Publishing, 2021.

This entry was readapted from *Guidotti, Monreale, Pedreschi, Giannotti. Principles of Explainable Artificial Intelligence. Springer International Publishing (2021)* by Francesca Pratesi and Riccardo Guidotti.

Safety and Robustness

In Brief

Safety and Robustness: The safety of an AI system refers to the extent the system meets its intended functionality without producing any physical or psychological harm, especially to human beings, and by extension to other material or immaterial elements that may be valuable for humans, including the system itself. Safety must also cover the way and conditions in which the system ceases its operation, and the consequences of stopping. The term robustness emphasises that safety and —conditionally to it— functionality, must be preserved under harsh conditions, including unanticipated errors, exceptional situations, unintended or intended damage, manipulation or catastrophic states.

Abstract

In this part we will cover the main elements that define the safety and robustness of AI systems. Some of them are common to system safety in general, to software-hardware computer systems or to critical systems engineering, such as **software bugs**. Some others are magnified in artificial intelligence, such as **denial of service**, a robustness issue that can appear by inducing an AI system to unrecoverable states or by generating inputs that collapse the system due to high computational demands. Some other issues are more specific to AI systems, such as **reward hacking**. These new issues appear more clearly in those systems that are specified in non-programmatic or non-explicit ways (e.g., through a utility function to be optimised, through examples, rewards or other implicit ways), as exemplified by systems that operate with solvers or machine learning models. We will pay more attention to these more AI-specific issues because they are less covered in the traditional literature about safety in computer systems. They are also more challenging because of their cognitive character, the ambiguities of human intent, several ethical issues and the relevance of long-term risks. This character and the fast development of the field has also blurred some distinctions between safety (threats without malicious intent) and Security (intentional threats), especially in now popular research areas such as Adversarial Attack and data_poisoning, and also within data privacy (e.g., **information leakage** by querying machine learning models or other **side channel attacks**). In the end, protecting the environment from the system (safety) also requires protecting the system from the environment (Security). Taking into account the changing character of the field, we include a taxonomic organisation of terms in the area of AI safety and robustness and their definition.

Motivation and Background

Given the increasing capabilities and widespread use of artificial intelligence, there is a growing concern about its risks, as humans are progressively replaced or sidelined from the decision loop of intelligent machines. The technical foundations and assumptions on which traditional safety engineering principles are based are inadequate for systems in which AI algorithms, and in particular Machine Learning (ML) algorithms, are interacting with people and the environment at increasingly higher levels of autonomy. There have been regulatory efforts to limit the use of AI systems in safety-critical or hostile environments, such as health, defense, energy, etc. [1, 2], but the consequences can also be devastating in areas that were not considered high risk, just by the scaling numbers or domino effects of AI systems. On top of the numerous safety challenges posed by present-day AI systems, a forward-looking analysis on more capable future AI systems raises more systemic concerns, such as highly disruptive scenarios in the workplace, the effect on human cognition in the long term and even existential risks.

Guidelines

Actions to ensure safety and robustness of AI systems need to take a holistic perspective, encompassing all the elements and stages associated with the conception, design, implementation and maintenance of these systems. We organise the field of AI safety and robustness into seven groups, following similar categorisations[1]:

- **AI Safety Foundations**: This category covers a number of foundational concepts, characteristics and problems related to AI safety that need special consideration from a theoretical perspective. This includes concepts such as uncertainty, generality or value alignment, as well as characteristics such autonomy levels, safety criticality, types of human-machine and environment-machine interaction. This group intends to collect any cross-category concerns in AI Safety and Robustness.
- **Specification and Modelling**: The main scope of this category is on how to describe needs, designs and actual operating AI systems from different perspectives (technical concerns) and abstraction levels. This includes the specification and modelling of risk management properties (e.g., hazards, failures modes, mitigation measures), as well as safety-related requirements, training, behaviour or quality attributes in AI-based systems.
- **Verification and Validation**: This category concerns design and implementation-time approaches to ensure that an AI-based system meets its requirements (verification) and behaves as expected (validation). The range of techniques covers any formal/mathematical, model-based simulation or testing approach that provides evidence that an AI-based system satisfies its defined (safety) requirements and does not deviate from its intended behaviour and causes unintended consequences, even in extreme and unanticipated situations (robustness).
- **Runtime Monitoring and Enforcement**: The increasing autonomy and learning nature of AI-based systems is particularly challenging for their verification and validation (V&V), due to our inability to collect an epistemologically sufficient quantity of evidence to ensure correctness. Runtime monitoring is useful to cover the gaps of design-time V&V by observing the internal states of a given system and its interactions with external entities, with the aim of determining system behaviour correctness or predicting potential risks. Enforcement deals with runtime mechanisms to self-adapt, optimise or reconfigure system behaviour with the aim of supporting fallback to a safe system state from the (anomalous) current state.
- **Human-Machine Interaction**: As autonomy progressively substitutes cognitive human tasks, some kind of human-machine interaction issues become more critical, such as the loss of situational awareness or overconfidence. Other issues include: collaborative missions that need unambiguous communication to manage self-initiative to start or transfer tasks; safety-critical situations in which earning and maintaining trust is essential at operational phases; or cooperative human-machine decision tasks where understanding machine decisions are crucial to validate safe autonomous actions.
- **Process Assurance and Certification**: Process Assurance is the planned and systematic activities that assure system lifecycle processes conform to its requirements (including safety) and quality procedures. In our context, it covers the management of the different phases of AI-based systems, including training and operational phases, the traceability of data and artefacts, and people. Certification implies a (legal) recognition that a system or process complies with industry standards and regulations to ensure it delivers its intended functions safely. Certification is challenged by the inscrutability of AI-based systems and the inability to ensure functional safety under uncertain and exceptional situations prior to its operation.

- **Safety-related Ethics, Security and Privacy**: While these are quite large fields, we are interested in their intersection and dependencies with safety and robustness. Ethics becomes increasingly important as autonomy (with learning and adaptive abilities) involves the transfer of safety risks, responsibility, and liability, among others. AI-specific security and privacy issues must be considered with regard to its impact on safety and robustness. For example, malicious adversarial attacks can be studied with focus on situations that compromise systems towards a dangerous situation.

Fig. 13 reflects the seven categories described above. Many of the terms and concepts we will expand on correspond to one or more of these categories.

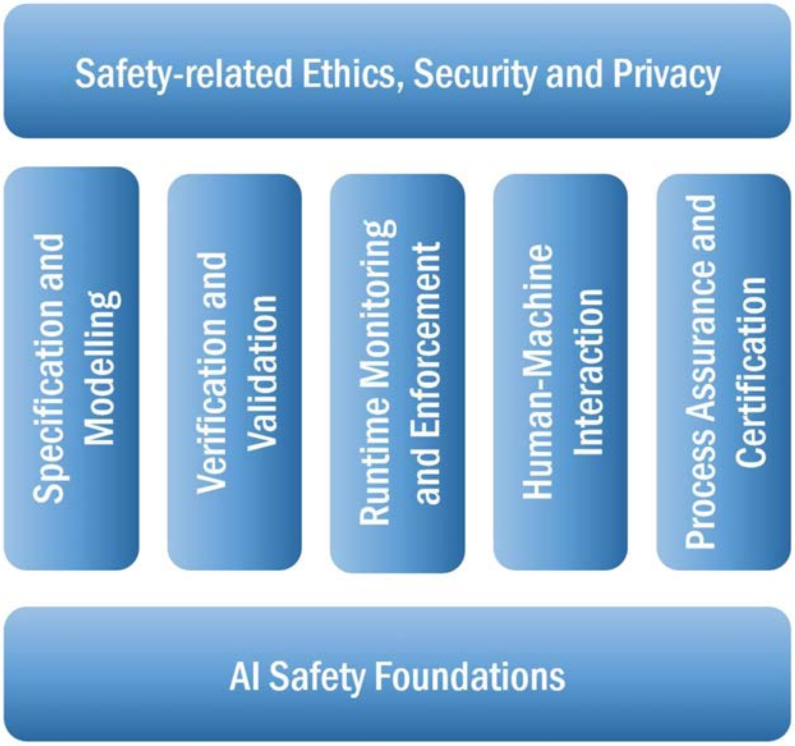

Safety-related Ethics, Security and Privacy
Specification and Modelling
Verification and Validation
Runtime Monitoring and Enforcement
Human-Machine Interaction
Process Assurance and Certification
AI Safety Foundations

Fig. 13 Taxonomy of AI Safety. Taken from [3]-

Main Keywords

- Alignment: The goal of AI **alignment** is to ensure that AI systems are aligned with human intentions and values. This first requires determining the normative question of what values or principles we have and what humans really want, collectively or individually, and second, the technical question of how to imbue AI systems with these values and goals..
- Robustness: **Robustness** is the degree in which an AI system functions1 reliably and accurately under harsh conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is therefore the strength of a system's integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, and undesirable reinforcement learning behaviour.
- Reliability: The objective of reliability is that an AI system behaves exactly as its designers intended and anticipated, over time. A reliable system adheres to the specifications it was programmed to carry out at any time. Reliability is therefore a measure of consistency of operation and can establish confidence in the safety of a system based upon the dependability with which it operationally conforms to its intended functionality.
- Evaluation: **AI measurement** is any activity that estimates attributes as *measures*— of an AI system or some of its components, abstractly or in particular contexts of operation. These attributes, if well estimated, can be used to explain and predict the *behaviour* of the system. This can stem from an engineering perspective, trying to understand whether a particular AI system meets the specifications or the intention of their designers, known respectively as **verification** and **validation**. Under this perspective, AI

measurement is close to computer systems **testing** (hardware and/or software) and other evaluation procedures in engineering. However, in AI there is an extremely complex *adaptive* behaviour, and in many cases, with a lack of a written and operational specification. What the systems has to do depends on some constraints and utility functions that have to be optimised, is specified by example (from which the system has to learn a model) or ultimately depends on feedback from the user or the environment (e.g., in the form of rewards).

- Negative side effects: **Negative side effects** are an important safety issue in AI system that considers all possible unintended harm that is caused as a secondary effect of the AI system's operation. An agent can disrupt or break other systems around, or damage third parties, including humans, or can exhaust resources, or a combination of all this. This usually happens because many things the system should *not* do are not included in its specification. In the case of AI systems, this is even more poignant as written specifications are usually replaced by an optimisation or loss function, in which it is even more difficult to express these things the system should not do, as they frequently rely on 'common sense'.
- Distributional shift: Once trained, most machine learning systems operate on static models of the world that have been built from historical data which have become fixed in the systems' parameters. This freezing of the model before it is released 'into the wild' makes its accuracy and reliability especially vulnerable to changes in the underlying distribution of data. When the historical data that have crystallised into the trained model's architecture cease to reflect the population concerned, the model's mapping function will no longer be able to accurately and reliably transform its inputs into its target output values. These systems can quickly become prone to error in unexpected and harmful ways. In all cases, the system and the operators must remain vigilant to the potentially rapid concept drifts that may occur in the complex, dynamic, and evolving environments in which your AI project will intervene. Remaining aware of these transformations in the data is crucial for safe AI.
- Security: The goal of **security** encompasses the protection of several operational dimensions of an AI system when confronted with possible attacks, trying to take control of the system or having access to design, operational or personal information. A secure system is capable of maintaining the integrity of the information that constitutes it. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also keeps confidential and private information protected even under hostile or adversarial conditions.
- Adversarial Attack: An **adversarial** input is any perturbation of the input features or observations of a system (sometimes imperceptible to both humans and the own system) that makes the system fail or take the system to a dangerous state. A prototypical case of an adversarial situation happens with machine learning models, when an external agent maliciously modify input data –often in imperceptible ways– to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a 'stop' sign as a 'speed limit' sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection. A different but related type of adversarial attack is called Data Poisoning, but this involves a malicious compromise of data sources (used for training or testing) at the point of collection and pre-processing.
- Data Poisoning: **Data poisoning** occurs when an adversary modifies or manipulates part of the dataset upon which a model will be trained, validated, or tested. By altering a selected subset of training inputs, a poisoning attack can induce a trained AI system into curated misclassification, systemic malfunction, and poor performance. An especially concerning dimension of targeted data poisoning is that an adversary may introduce a 'backdoor' into the infected model whereby the trained system functions normally until it processes maliciously selected inputs that trigger error or failure. Data poisoning is possible because data collection and procurement often involves potentially unreliable or questionable sources. When data originates in uncontrollable environments like the internet, social media, or the Internet of Things, many opportunities present themselves to ill-intentioned attackers, who aim to manipulate training examples. Likewise, in third-party data curation processes (such as 'crowdsourced' labelling, annotation, and content identification), attackers may simply handcraft malicious inputs.

Recommended reading

Some introductory sources for AI Safety and Robustnes are [3, 1, 1, 1, 7].

Bibliography

[[1]] Luciano Floridi. The european legislation on ai: a brief analysis of its philosophical approach. *Philosophy & Technology*, 34(2):215–222, 2021.

[[2]] Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021.

[3](1,2) Huáscar Espinoza, Han Yu, Xiaowei Huang, Freddy Lecue, José Hernández-Orallo, Seán Ó hÉigeartaigh, and Richard Mallah. Towards an AI safety landscape: an overview. https://www.ai-safety.org/, 2019.

[[4]] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. 2016. arXiv:1606.06565.

[[5]] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[[6]] Leslie David. Understanding artificial intelligence ethics and safety. *The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529*, 2019.

[[7]] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114, 2015.

This entry was readapted from *Huáscar Espinoza, Han Yu, Xiaowei Huang, Freddy Lecue, José Hernández-Orallo, Seán Ó hÉigeartaigh, and Richard Mallah. Towards an AI safety landscape: an overview. Artificial Intelligence Safety 2019, https://www.ai-safety.org/.* by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

[[1]] FLI’s Landscape of AI Safety and Beneficence Research for research contextualization and in preparation for brainstorming at the Beneficial AI 2017 conference (https://futureoflife.org/landscape/ResearchLandscapeExtended.pdf), the Assuring Autonomy International Programme (AAIP) to develop a Body of Knowledge (BoK) intended, in time, to become a reference source on assurance and regulation of Robotics and Autonomous Systems (RAS), (https://www.york.ac.uk/assuring-autonomy/research/body-of-knowledge/) and Ortega et al (DeepMind) structure of the technical AI safety field (https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1).

Alignment

Synonyms: (Mis)directed, (Un)intended behaviour

In brief

The goal of AI **alignment** is to ensure that AI systems are aligned with human intentions and values. This first requires determining the normative question of what values or principles we have and what humans really want, collectively or individually, and second, the technical question of how to imbue AI systems with these values and goals.

More in detail

The concept of **alignment** has been mostly covered at the philosophical and ethical levels, because the normative question involves fundamental issues about human behaviour and ethics, and goes beyond the related concept of **validity**. Also, the technical question is hard to solve even if the normative question is clear, because it depends on a very diverse collection of paradigms about what an AI system is and what it is expected to be, in terms of techniques and capabilities.

Let us start with the normative question. Following Gabriel (2020) [1], there is no consensus of what alignment means:

> “there are significant differences between AI that aligns with instructions, intentions, revealed preferences, ideal preferences, interests and values”.

In particular, these six different perspectives can be summarised as follows:

- “Instructions: the agent does what I instruct it to do”.
- “Expressed intentions: the agent does what I intend it to do.”
- “Revealed preferences: the agent does what my behaviour reveals I prefer.”
- “Informed preferences or desires: the agent does what I would want it to do if I were rational and informed.”
- “Interest or well-being: the agent does what is in my interest, or what is best for me, objectively speaking.”
- “Values: the agent does what it morally ought to do, as defined by the individual or society.”

None of these interpretations fully captures what alignment should be, and some of them may lead to important problems and paradoxes. As said before, even in those cases where there could be some agreement and disambiguation in clear cases, the technical question is also fraught with difficulties. One general technical problem of aligning AI systems is that it is hard to say what the system has to do, but it is much harder to specify what the system should not do, mostly because this is taken for granted or appeals to “common” sense, which machines lack today.

Then, many specific problems manifest differently depending on the particular AI paradigm. For instance, in reinforcement learning (RL), “the learner system actively solves problems by engaging with its environment through trial and error. This exploration and ‘problem-solving’ behaviour is determined by the objective of maximising a reward function that is defined by its designers. […] An RL system, which is operating in the real-world without sufficient controls, may determine a reward-optimising course of action that is optimal for achieving its desired objective but harmful to people.” [1]. A significant research and philosophical effort about the AI systems of the future has been framed as a RL problem.

Other kinds of systems show different problems. For instance, digital assistants are commanded by natural language. As a result, alignment problems can come from misunderstanding of the commands, given the ambiguity of natural language. For instance, after the command “prepare something proteic for dinner” a digital robotic assistant may put the cat in the oven. Even non-agential systems such as a simple supervised model may end up terribly misaligned with human values by being trained by biased or narrow datasets. For instance, a self-driving car with a pedestrian recognition system may fail to detect a group of people being disguised at a carnival. These are examples of three kinds of AI systems (*reinforcement learning*, *digital assistants*, and *object recognition systems*) that have been around for some time, and we still face many safety and robustness issues with them. The problems with new paradigms, such as language models, are still being recognised.

Overall, AI alignment is a critical and fundamental open problem that requires philosophical, ethical and technical progress. The progress so far is not keeping up with the developments of AI as a field.

Bibliography

[[1]] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[[2]] Leslie David. Understanding artificial intelligence ethics and safety. *The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529*, 2019.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

Robustness

Synonyms: Brittleness.

In brief

Robustness is the degree in which an AI system functions[1] reliably and accurately *under harsh conditions.* These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is therefore the strength of a system's integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, and undesirable reinforcement learning behaviour.

More in detail

Robustness is a broad term that usually encompasses many problems that can affect the stability and good behaviour of an AI system. However, it does not cover the failure of a system under normal operation, attrition or obsolescence, issues that are covered by the related term of Reliability. Robustness is also closely related to Security, as in both cases the system must withstand (adversarial) attacks. However, robustness does not usually cover elements such as unauthorised access that compromises privacy, but only those that can lead to operational failure or damage.

Robustness can be ensured by *prevention* or by *recovery* procedures. The prevention aspect of robustness has to do with preventive testing of the functioning of the AI System under uncommon or stressful conditions. This has been argued by the European Commission as a key piece to ensure the trustworthiness of AI systems [2], and is similar to the testing any software would undergo before taking decisions in the real world, from aircraft control systems to banking web pages. Testing for robustness not only considers attrition over time, covered by Reliability, but especially in abnormal working mode, such as for example when human users make mistakes. An example of this is the automatic brake system cars introduce: in normal conditions the human will be in charge of braking. However, since the car is designed with collision prevention in mind, it should also be robust to human errors.

The *recovery* procedure, on the other hand, ensures that even if the AI system is not able to prevent the failure, it will limit the amount of damage produced. For example, if a conversational AI support system is unsure how to respond to specific queries, it may still have the safe policy of deferring to a human agent.

Robustness is compromised when systems are brittle to unfamiliar events and scenarios. They may make unexpected and serious mistakes, because they have neither the capacity to contextualise the problems they are programmed to solve nor the common-sense ability to determine the relevance of new 'unknowns'. This fragility or brittleness can have especially significant consequences in safety-critical applications like fully automated transportation and medical decision support systems where undetectable changes in inputs may lead to significant failures. Alternatively, **robustness** might also be critical in situations where it is very hard for a human to intervene and manually recover from the error, such as for instance, in a space mission.

Bibliography

[[1]] Leslie David. Understanding artificial intelligence ethics and safety. *The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529*, 2019.

[[2]] European Commission. On artificial intelligence—a european approach to excellence and trust. 2020.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

[1] From here the definition is taken from [1] under Creative Commons Attribution License 4.0.

Reliability

Synonyms: Dependability.

In brief

The objective of **reliability** [1] is that an AI system behaves exactly as its designers intended and anticipated, *over time*. A reliable system adheres to the specifications it was programmed to carry out at any time. Reliability is therefore a measure of consistency of operation and can establish confidence in the safety of a system based upon the dependability with which it operationally conforms to its intended functionality.

More in detail

The usual definition of **reliability** is the probability of a system performing its intended functions under expected conditions. Reliability is closely related to Robustness and resilience, but the focus is on the time dimension, which is different from other safety considerations. That is, the system can perform its designed functionality for the intended period of time. Hardware reliability is in general well studied, or there are mature methods for testing and assessing hardware reliability. Thus, the focus of AI reliability, different from traditional reliability studies, is on the software system. Compared to hardware reliability, software reliability is typically more difficult to test, which brings challenges to the research and development of reliable AI systems.

Kaur and Bahl [1] defined the reliability of software as “the probability of the failure-free software operation for a specified period of time in a specified environment”. There are thus three key elements in the definition of reliability, “failure”, “time” and “environment” (or “operational profile”). Failure means that in some way the software has not functioned according to the customer’s requirements [2]. The failure events of an AI system can be mostly related to software errors, in addition to the failure of hardware. For hardware failures, [3] discussed that AI hardware failures are related to software errors, aging, process variation, and temperature. Software failures, understood as a departure of the external behavior of the program from the user’s requirements, are usually related to software errors and interruptions. For example, the occurrence of a disengagement event is considered as a failure of the system for autonomous vehicles [4]). Note also the related term of software “fault”, which refers to a defect in a program that, when executed under certain conditions, causes a failure—that is, what is generally called a “bug”. The time scale in AI reliability can be different for different structure levels or AI applications (e.g., calendar time, cycles, calls, etc., to AI algorithms, or, miles driven for autonomous vehicles or the length of a conversation between a customer and an AI chatbot, when analysing particular applications). Finally, the operating environment includes both the physical environment for hardware systems (e.g., compute, temperature, humidity, etc.), and non-physical for software systems (e.g., data, libraries, meta-settings, etc.).

Software reliability is not only one of the most important and immediate attributes of software quality, it is also the most readily quantified and measured. From the basic notion of reliability, many different measures can be developed to quantify the occurrence of failures in time. Some of the most important of these measures, and their interrelationships are summarised below (adapted from [2]):

- **Hazard Rate**, denoted by $z(t)$, is the conditional failure density at time t, given that no failure has occurred up to that time. That is, $z(t) = f(t)/R(t)$, where $f(t)$, is the probability density for failure at time t, and $R(t)$ is the probability of failure-free operation up to time t. Reliability and hazard rate are related by $R(t) = e^{-\int_{0}^{t} z(x) dx}$.
- **Mean Value Function**, denoted by $\mu(t)$, is the mean number of failures that have occurred by time t.
- **Failure Intensity**, denoted by $\lambda(t)$, is the number of failures occurring per unit time at time t. This is related to the mean value function by $\lambda(t) = \frac{d}{dt} \mu(t)$. The number of failures expected to occur in the half open interval $(t,t + \delta t]$ is $\lambda(t) \dot \delta t$.

Failure intensity is the measure most commonly used in the quantification of software reliability [2]. Software reliability models have appeared as people try to understand the features of how and why software fails, and attempt to quantify software reliability. Given that the quantities associated with reliability are usually random variables (due to of the complexity of the factors influencing the occurrence of a failure), reliability models follow the form of random stochastic processes defining the behaviour of software failures to time. Over 200 models have been established since the early 1970s (see [5] for a survey of software reliability models). Since the number of faults in a program generally changes over time (as they are usually repaired when they appear), the probability distributions of the components of a reliability model vary with time and, thus, reliability models are based on nonhomogeneous random processes.

Finally, it should be stressed that, when people desire extremely high reliability because of the critical nature of a particular application, e.g. for autopilot software, they often use formal logical systems to maximise their certainty of implementation correctness [6, 7].

While many of the concepts in software reliability apply to artificial intelligence, some approaches are not directly applicable because the lack of a clear specification, and accordingly there are many other sources of faults that are not software 'bugs' or hardware errors.

Bibliography

[[1]] Gurpreet Kaur and Kailash Bahl. Software reliability, metrics, reliability improvement using agile process. *International Journal of Innovative Science, Engineering & Technology*, 1(3):143–147, 2014.

[2](1,2,3) John Rushby. *Quality measures and assurance for AI software*. Volume 18. National Aeronautics and Space Administration, Scientific and Technical …, 1988.

[[3]] Muhammad Abdullah Hanif, Faiq Khalid, Rachmad Vidya Wicaksana Putra, Semeen Rehman, and Muhammad Shafique. Robust machine learning systems: reliability and security for deep neural networks. In *2018 IEEE 24th International Symposium on On-Line Testing And Robust System Design (IOLTS)*, 257–260. IEEE, 2018.

[[4]] Yili Hong, Jie Min, Caleb B King, and William Q Meeker. Reliability analysis of artificial intelligence systems using recurrent events data from autonomous vehicles. *arXiv preprint arXiv:2102.01740*, 2021.

[[5]] John D Musa, Anthony Iannino, and Kazuhira Okumoto. Software reliability. *Advances in computers*, 30:85–170, 1990.

[[6]] Donald C Latham. Department of defense trusted computer system evaluation criteria. *Department of Defense*, 1986.

[[7]] Stuart Russell. Unifying logic and probability. *Communications of the ACM*, 58(7):88–97, 2015.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

[1] Definition taken from {cite}`david2019understanding under Creative Commons Attribution License 4.0.

Evaluation

Synonyms: Assessment, Testing, Measurement.

In brief

AI measurement is any activity that estimates attributes as *measures*— of an AI system or some of its components, abstractly or in particular contexts of operation. These attributes, if well estimated, can be used to explain and predict the *behaviour* of the system. This can stem from an engineering perspective, trying to understand whether a particular AI system meets the specifications or the intention of their designers, known respectively as **verification** and **validation**. Under this perspective, AI measurement is close to computer systems **testing** (hardware and/or software) and other evaluation procedures in engineering. However, in AI there is an extremely complex *adaptive* behaviour, and in many cases, with a lack of a written and operational specification. What the systems has to do depends on some constraints and utility functions that have to be optimised, is specified by example (from which the system has to learn a model) or ultimately depends on feedback from the user or the environment (e.g., in the form of rewards).

More in detail

AI **measurement** has been taken place since the early days of AI and has framed the discipline very significantly. Actually, one of the foundational ideas behind AI is the famous imitation game [2], which —somewhat misleadingly— is usually referred to as the Turing *test*. However, this initial emphasis on evaluation, albeit mostly philosophical, did not develop into technical AI evaluation as an established subfield in AI, in the same way the early Reliability and Robustness problems in software engineering led to the important areas of software validation, verification and testing, today using both theoretical and experimental approaches [5, 3]. Still, there are some recent surveys and books covering the problem of AI evaluation, and giving comprehensive or partial views of AI measurement, such as [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

The tradition in AI measurement turns around the concept of 'task-oriented evaluation'. For instance, given a scheduling problem, a board game or a classification problem, systems are evaluated according to some metric of *task performance*. To standardise the comparisons among systems, there are datasets and benchmarks that are used for evaluating these systems, so that evaluation data is not cherry-picked by the AI designers. We find a myriad of examples of the latter in PapersWithCode[1] (PwC), an open-source, community-centric platform which offers researchers access to hundreds of benchmarks and thousands of results from associated papers, with an emphasis on machine learning. PwC collects information about the performance of different AI systems during a given period, typically ranging from the introduction of the benchmark to the present day. Figure 14 shows an example of the evolution of results for ImageNet[15].

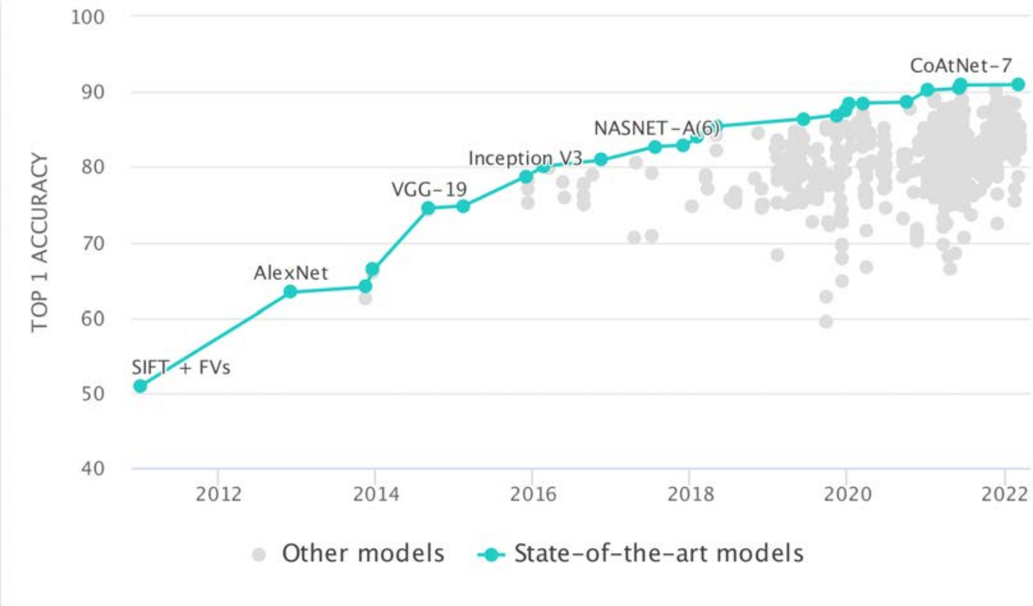


Fig. 14 State-of-the-Art for the image classification task using the benchmark ImageNet. The points represent the accuracy of all the attempts from 2011 to 2022. The connected points on the Pareto front are shown in blue. Chart from https://paperswithcode.com/sota/image-classification-on-imagenet

However, the focus on particular benchmarks (which are known by the researchers in advance) and the power of machine learning techniques has led to a problem of benchmark specialisation, a phenomenon that is related to issues such as "teaching to the test" (students prepare for the test but do not know how to solve cases that differ slightly from those of the exam) or Goodhart's or Campbell's laws (optimising to the indicator may lead to the metric not measuring what it measured originally, with possibly some other side effects). One clear manifestation of this phenomenon is the 'challenge-solve-and-replace' evaluation dynamics [16] or a 'dataset-solve-and-patch' adversarial benchmark co-evolution [17], which means that as soon as a benchmark is released, performance grows quickly because researchers specialise the design and training of the system to the benchmark, but not to the general task [11]. Ultimately the benchmark needs to be replaced by another one (usually more complex or adversarially designed), in a continuous cycle.

This task-oriented evaluation has been blamed for some of the failures or narrowness of AI in the past —lack of common sense, of generality, of adaptability to new contexts and distributions. As we said, since the beginning of the discipline, other approaches for AI measurement have been used or proposed. These include the Turing test, and endless variants [18, 19], the use of human tests, from science exams [20] to psychometric tests [21], the adaptation of psychophysics [22] or item response theory [23], the use of collections of video games [24], the exploration of naive physics [25], or the adaptation of tests from animal cognition [26]. All these approaches attempt to measure intelligence more broadly, some general cognitive abilities or at least skills that could be applied to a range of different tasks. Accordingly, these fall under the paradigm of ‘capability-oriented evaluation’ [4].

The key difference between performance and capability is that performance is affected by the distribution, while capability is not. For instance, the same individual (an AI system or a human) can have different degrees of performance for the same task or set of tasks if we change the distribution of examples (e.g., by including more difficult examples), but the capability should be the same, since it should be a property of an individual. If a person or computer has the capability of resolving simple negation, this *capability* is not changed by including many double negations in the dataset, even if this decreases performance. However, identifying and estimating the level for different capabilities is much more challenging than measuring performance. Also, drawing conclusions about the cognitive abilities of AI systems requires caution, even from the most-well designed experiments. But this is also true even when performance is used as a proxy for capability [27].

Bibliography

[[1]] John D Musa, Anthony Iannino, and Kazuhira Okumoto. Software reliability. *Advances in computers*, 30:85–170, 1990.

[[2]] A.M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433, 1950.

[[3]] W Richards Adrion, Martha A Branstad, and John C Cherniavsky. Validation, verification, and testing of computer software. *ACM Computing Surveys (CSUR)*, 14(2):159–192, 1982.

[4](1,2) José Hernández-Orallo. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, 48(3):397–447, 2017.

[[5]] J. Hernández-Orallo. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press, 2017.

[[6]] José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, and Kristinn R Thórisson. A new AI evaluation cosmos: ready to play the game? *AI Magazine*, 2017.

[[7]] Chris Welty, Praveen Paritosh, and Lora Aroyo. Metrology for AI: from benchmarks to instruments. *arXiv preprint arXiv:1911.01875*, 2019.

[[8]] Peter Flach. Performance evaluation in machine learning: the good, the bad, the ugly and the way forward. In *AAAI*. 2019.

[[9]] Peter Flach. Measurement theory for data science and AI: modelling the skills of learning machines and developing standardised benchmark tests. Turing Institute, https://www.turing.ac.uk/research/research-projects/measurement-theory-data-science-and-ai, 2019.

[[10]] Guillaume Avrin. Evaluation of artificial intelligence systems. Laboratoire National de Métrologie et d’Essais : https://www.lne.fr/en/testing/evaluation-artificial-intelligence-systems, 2019.

[11](1,2) Jose Hernandez-Orallo. Ai evaluation: on broken yardsticks and measurement scales. In *AAAI Workshop on Evaluating Evaluation of AI Systems*. 2020.

[[12]] Fernando Martínez-Plumed and José Hernández-Orallo. Dual indicators to analyze ai benchmarks: difficulty, discrimination, ability, and generality. *IEEE Transactions on Games*, 12(2):121–131, 2020. doi:10.1109/TG.2018.2883773.

[[13]] **missing journal in hernandez2021identifying**

[[14]] Jose Hernandez-Orallo, Wout Schellaert, and Fernando Martinez-Plumed. Training on the test set: mapping the system-problem space in ai. *AAAI Senior Member Track - Blue Sky Ideas*, 2022.

[[15]] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee, 2009.

[[16]] David Schlangen. Language tasks and language games: on methodology in current natural language processing research. *arXiv preprint arXiv:1908.10747*, 2019.

[[17]] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[[18]] G. Oppy and D. L. Dowe. The Turing Test. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. 2011. Stanford University, http://plato.stanford.edu/entries/turing-test/.

[[19]] José Hernández-Orallo. Twenty years beyond the turing test: moving beyond the human judges too. *Minds and Machines*, 30(4):533–562, 2020.

[[20]] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[[21]] S. Bringsjord and B. Schimanski. What is artificial intelligence? Psychometric AI as an answer. In *International Joint Conference on Artificial Intelligence*, 887–893. 2003.

[[22]] Joel Z Leibo and others. Psychlab: a psychology laboratory for deep reinforcement learning agents. *arXiv preprint arXiv:1801.08116*, 2018.

[[23]] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in AI: analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019.

[[24]] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: an evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[[25]] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: a new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 2019.

[[26]] Benjamin Beyret, José Hernández-Orallo, Lucy Cheke, Marta Halina, Murray Shanahan, and Matthew Crosby. The animal-ai environment: training and testing animal-like artificial cognition. *arXiv preprint arXiv:1909.07483*, 2019.

[[27]] Ryan Burnell, John Burden, Danaja Rutar, Konstantinos Voudouris, Lucy Cheke, and José Hernández-Orallo. Not a number: identifying instance features for capability-oriented evaluation. *IJCAI*, 2022.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

[[1]] paperswithcode.com

Negative side effects

In brief

Negative side effects are an important safety issue in AI system that considers all possible unintended harm that is caused as a secondary effect of the AI system’s operation. An agent can disrupt or break other systems around, or damage third parties, including humans, or can exhaust resources, or a combination of all this. This usually happens because many things the system should *not* do are not included in its specification. In the case of AI systems, this is even more poignant as written specifications are usually replaced by an optimisation or loss function, in which it is even more difficult to express these things the system should not do, as they frequently rely on ‘common sense’.

More in detail

Negative Side Effects are unanticipated or unintended effects caused by an AI system during operation. Negative side effects are a key issue within the AI safety literature [1] and typically stem from the difficulty of fully articulating everything that we want the AI system *not* to do. Negative side effects can be of two kinds, related to some preservation of the environment or related to the use of resources. For instance, a clumsy incompetent agent may break everything around it [2], not preserving things that should be unrelated to the goal. On the other hand, a very proficient agent may exhaust all available resources and disrupt any other agent (machine or human) that interferes with its goals [3]. Recognising these two different causes for side effects is crucial for the design of safer AI systems.

There are a number of proposed methods for measuring the negative side effects caused by a system [4, 5, 6], but a recurring theme is based on estimating counterfactual scenarios in which the system was not present (or acted differently), and comparing the changes in the state of the world. Alternatively, in some approaches for mitigating side effects, these are set externally (marked safe areas) or incorporated as a “secondary objective” [7] for which trade-offs are found. But in these approaches side effects just become part of the specification or the optimisation function, significantly deviating from the original definition of *side* effect.

For machine learning systems, the real challenge of side effects is that many of them are not known during training, and they cannot be incorporated as a regulariser in the target function to be optimised. In many cases, this information is not even available to the agent during operation, so we cannot modify the agent’s behaviour according to this information, such as marking some forbidden areas of operation.

While some other safety issues in RL have received important attention [8], dealing with side effects is still mostly unexplored, especially in realistic situations where the agent does not have any feedback about side effects during training or operation. Also, the analysis of AI safety in reinforcement learning has usually been conducted with toy scenarios, not having the complex interaction of the real world. SafeLife [9] is an exception in this landscape, as a very rich and full *ecosystem* where many complex behaviours and effects can be analysed.

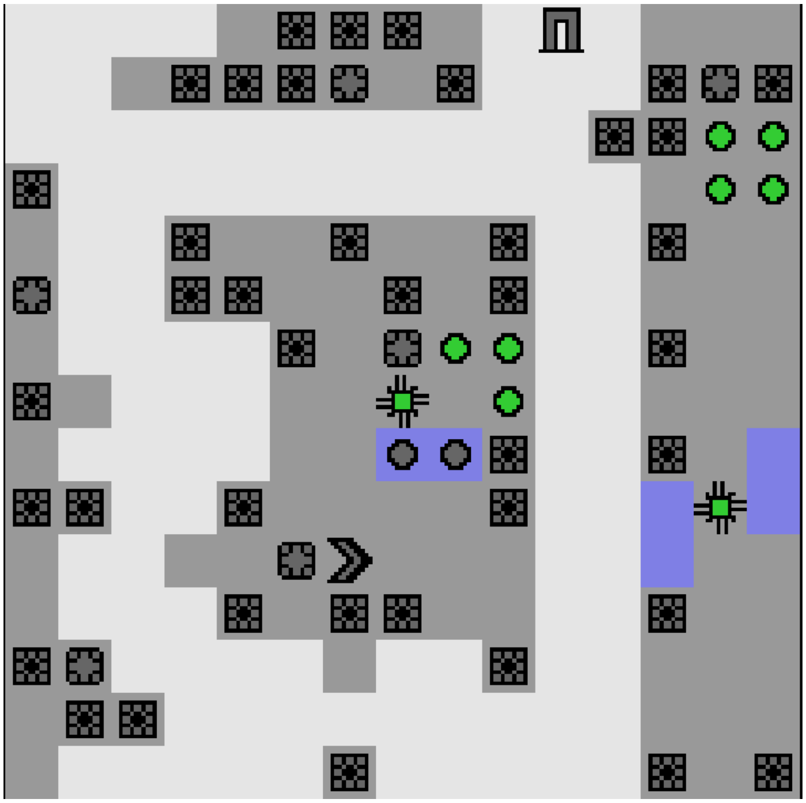


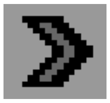
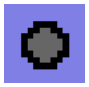
Fig. 15 An example task instance from SafeLife. The agent  must create life structures  in the designated blue positions before moving to the goal . Ideally the agent should not disturb the green life cells .

Fig. 15 gives the visual representation of a state of a SafeLife instance. There we can see that achieving the goals can be done carefully so that the green life cells are not disturbed. The real challenge is when the system is not informed explicitly (as part of the reward or a constraint) that the green life cells should not be disturbed. In this case, only agents that try to be minimise changes that are not necessary for the goals will be expected to have low side effects.

Bibliography

[[1]] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. 2016. arXiv:1606.06565.

[[2]] Sandhya Saisubramanian, Shlomo Zilberstein, and Ece Kamar. Avoiding negative side effects due to incomplete knowledge of ai systems. *arXiv preprint arXiv:2008.12146*, 2020.

[[3]] Roman V Yampolskiy. *Artificial intelligence safety and security*. CRC Press, 2018.

[[4]] Stuart Armstrong and Benjamin Levinstein. Low impact artificial intelligences. *arXiv preprint arXiv:1705.10720*, 2017.

[[5]] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. 2017. arXiv:1711.09883.

[[6]] Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Feb 2020.

[[7]] Sandhya Saisubramanian, Ece Kamar, and Shlomo Zilberstein. A multi-objective approach to mitigate negative side effects. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*. 2020.

[[8]] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[[9]] Carroll L. Wainwright and Peter Eckersley. Safelife 1.0: exploring side effects in complex environments. 2019. arXiv:1912.01217.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

Distributional shift

Synonyms: Data shift.

In brief

Once trained, most machine learning systems operate on static models of the world that have been built from historical data which have become fixed in the systems’ parameters. This freezing of the model before it is released ‘into the wild’ makes its accuracy and reliability especially vulnerable to changes in the underlying distribution of data. When the historical data that have crystallised into the trained model’s architecture cease to reflect the population concerned, the model's mapping function will no longer be able to accurately and reliably transform its inputs into its target output values. These systems can quickly become prone to error in unexpected and harmful ways. In all cases, the system and the operators must remain vigilant to the potentially rapid concept drifts that may occur in the complex, dynamic, and evolving environments in which your AI project will intervene. Remaining aware of these transformations in the data is crucial for safe AI. [1]

More in detail

A common use case of machine learning in real world settings is to learn a model from historical data and then deploy the model on future unseen examples. When the data distribution for the future examples differs from the historical data distribution (i.e., the joint distribution of inputs and outputs differs between training and test o deployment stages), machine learning techniques that depend precariously on the i.i.d. assumption tend to fail. This phenomena is call distributional shift and is a very common problem [3]. Note that a particular case of distributional shift occurs when only the input distribution changes (covariate shift) or there is a shift in the target variable (prior probability shift).

The problem of distributional shift is of relevance not only to academic researchers, but to the machine learning community at large. Distributional shift is present in most practical applications, for reasons ranging from the bias introduced by experimental design to the irreproducibility of the testing conditions at training time. An example is email spam filtering, which may fail to recognise spam that differs in form from the spam the automatic filter has been built on [4], yet often the model being highly confident in its erroneous classifications. This issue is especially important in high-risk applications of machine learning, such as finance, medicine, and autonomous vehicles, where a mistake may incur financial or reputational loss, or possible loss of life. It is therefore important to assess both a model's robustness to distribution shift and its estimates of predictive uncertainty, which enable it to detect distributional shifts [1, 5].

In general, the greater the degree of shift, the poorer the model's performance is. The performance of learned models tend to drop significantly even with a tiny amount of distribution shift between training and test [6, 7], which makes it challenging to reliably deploy machine learning in real world applications. Although one can always increase training coverage by adding more sources of data [8], data augmentation [9, 10], or injecting structural bias into models [11, 12, 13] for generalisation to any potential input for the learned model, it is unrealistic to expect a learned model to predict accurately under any form of distribution shift due to the combinatorial nature of real world data and tasks.

On the other hand, adapting a model to a specific type of distribution shift might be more approachable than adapting to any potential distribution shift scenarios, under appropriate assumptions and with appropriate modifications. By knowing where the model can predict well, one can use the model to make conservative predictions or decisions, and to guide future active data collection to covered shifted distributions. Therefore, in addition to improving the generalisation performance of models in general, methods that explicitly deal with the presence of distribution shift are also desirable for machine learning to be used in practice [14].

In terms of assessment, the robustness of learning models to distributional shift is typically assessed via metrics of predictive performance on a particular task: given two (or more) evaluation sets, where one is considered matched to the training data and the other(s) shifted, models which have a smaller degradation in performance on the shifted data are considered more robust. The quality of uncertainty estimates is often assessed via the ability to classify whether an example came from the "in-domain" dataset or a shifted dataset using measures of uncertainty.

For its part, concept shift (or concept drift) is different from distributional shift in that it is not related to the input data or the class distribution but instead is related to the relationship between two or more dependent variables. An example may be the customer purchasing behavior over time in a particular online shop. This behaviour may be influenced by the strength of the economy, this being not explicitly specified in the data. In this case, the concept of interest (consumer behaviour) depends on some hidden context, not known a priori, and not given explicitly in the form of predictive features, making the learning task more complicated [15]. In this sense, concept shift can be categorised into 3 types:

1. *sudden, abrupt or instantaneous concept shift* (e.g., following the previous example, the COVID-19 lockdowns significantly changed customer behaviour);

2. *gradual concept shift* (e.g., customers are influenced by wider economic issues, unemployment rates or other trends) which can be divided further into moderate and slow drifts, depending on the rate of the changes [16];
3. *cyclic concept drifts*, where hidden contexts may be expected to recur due to cyclic phenomena, such as seasons of the year or may be associated with irregular phenomena, such as inflation rates or market mood [17].

Concept drift may be present on supervised learning problems where predictions are made and data is collected over time. These are traditionally called online or incremental learning problems [18], given the change expected in the data over time. For its part, the common methods for detecting concept drift in machine learning generally include ongoing monitoring of the performance (e.g., accuracy) and confidence scores of a learning model. If average performance or confidence deteriorates over time, concept shift could be occurring

Bibliography

[[1]] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. 2016. arXiv:1606.06565.

[[2]] Leslie David. Understanding artificial intelligence ethics and safety. *The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529*, 2019.

[[3]] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[[4]] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 2006.

[[5]] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR, 2016.

[[6]] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. PMLR, 2019.

[[7]] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arxiv 2013. *arXiv preprint arXiv:1312.6199*, 2013.

[[8]] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[[9]] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[[10]] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[[11]] Kunihiko Fukushima and Sei Miyake. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[[12]] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[[13]] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.

[[14]] Yifan Wu. *Learning to Predict and Make Decisions under Distribution Shift*. PhD thesis, University of California, 2021.

[[15]] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.

[[16]] Kenneth O Stanley. Learning concept drift with a committee of decision trees. *Informe técnico: UT-AI-TR-03-302, Department of Computer Sciences, University of Texas at Austin, USA*, 2003.

[[17]] Michael Bonnell Harries, Claude Sammut, and Kim Horn. Extracting hidden context. *Machine learning*, 32(2):101–126, 1998.

[[18]] Gregory Ditzler and Robi Polikar. Incremental learning of concept drift from streaming imbalanced data. *IEEE transactions on knowledge and data engineering*, 25(10):2283–2301, 2012.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

[[1]] Definition taken from [1] under Creative Commons Attribution License 4.0.

Security

In brief

The goal of **security** encompasses the protection of several operational dimensions of an AI system when confronted with possible attacks, trying to take control of the system or having access to design, operational or personal information. A secure system is capable of maintaining the integrity of the information that constitutes it. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also keeps confidential and private information protected even under hostile or adversarial conditions. [1]

More in detail

Security must be an integral part of the AI process. Protecting AI systems, their data and their communications is critical to the security and privacy of users, as well as protecting business investments. The AI systems themselves are incredibly expensive and possess valuable intellectual property to protect against disclosure and misuse. The confidentiality of the program code associated with AI systems may be considered less critical, but access to it, as well as the ability to manipulate this code, can result in the disclosure of important and confidential assets.

Several kinds of attacks against AI systems have been reported. Currently, the most prominent attack vector categories are [2]: adversarial inputs [3]; data poisoning attacks [4]; model stealing techniques [5, 6]; model poisoning [7], data leakage [8] and neural network Trojans [9], among others. Attack vectors directed against the AI systems' deployment or training environment are equally applicable. These may be attack vectors directed against servers, databases, protocols or libraries utilised within the AI system [10].

Currently, AI systems often lack sufficient security assessments [11]. This may be the result of the mutually independent development of AI methods and their implementation in applications: while the application should have a security assessment, embedded AI (via APIs or frameworks) is rarely considered in terms of its security vulnerabilities by application developers and/or practitioners. While AI developers may follow coding standards and guidelines for secure software development, they will not assess the potential attack surface of an AI system (i.e., the means by which an attacker may enter, extract data or manipulate the system in question) using the system.

Bibliography

[[1]] Leslie David. Understanding artificial intelligence ethics and safety. *The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529*, 2019.

[[2]] Xiaofeng Liao, Liping Ding, and Yongji Wang. Secure machine learning, a brief overview. In *2011 Fifth International Conference on Secure Software Integration and Reliability Improvement-Companion*, 26–29. IEEE, 2011.

[[3]] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[[4]] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, 9389–9398. PMLR, 2021.

[[5]] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX security symposium (USENIX Security 16)*, 601–618. 2016.

[[6]] Raül Fabra-Boluda, Cèsar Ferri, José Hernández-Orallo, Fernando Martínez-Plumed, and María José Ramírez-Quintana. Identifying the machine learning family from black-box models. In *Conference of the Spanish Association for Artificial Intelligence*, 55–65. Springer, 2018.

[[7]] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to Byzantine-Robust federated learning. In *29th USENIX Security Symposium (USENIX Security 20)*, 1605–1622. 2020.

[[8]] Panagiotis Papadimitriou and Hector Garcia-Molina. Data leakage detection. *IEEE Transactions on knowledge and data engineering*, 23(1):51–63, 2010.

[[9]] Yu Ji, Zixin Liu, Xing Hu, Peiqi Wang, and Youhui Zhang. Programmable neural network trojan for pre-trained feature extractor. *arXiv preprint arXiv:1901.07766*, 2019.

[[10]] Kim Hartmann and Christoph Steup. Hacking the ai-the next generation of hijacked systems. In *2020 12th International Conference on Cyber Conflict (CyCon)*, volume 1300, 327–349. IEEE, 2020.

[[11]] Why ai needs security. https://www.synopsys.com/designware-ip/technical-bulletin/why-ai-needs-security-dwtb-q318.html, 2020.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

[[1]] Definition taken from [1] under Creative Commons Attribution License 4.0.

Adversarial Attack

Synonyms: Adversarial Input, Adversarial Example.

In brief

An **adversarial input** is any perturbation of the input features or observations of a system (sometimes imperceptible to both humans and the own system) that makes the system fail or take the system to a dangerous state. A prototypical case of an adversarial situation happens with machine learning models, when an external agent maliciously modify input data –often in imperceptible ways– to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a ‘stop’ sign as a ‘speed limit’ sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection. A different but related type of adversarial attack is called Data Poisoning, but this involves a malicious compromise of data sources (used for training or testing) at the point of collection and pre-processing.

More in detail

The vulnerabilities of AI systems to adversarial examples have serious consequences for AI safety. The existence of cases where subtle but targeted perturbations cause models to be misled into gross miscalculation and incorrect decisions have potentially serious safety implication for the adoption of critical systems like applications in autonomous transportation, medical imaging, and security and surveillance.

To get an idea of what adversarial examples look like, consider the example in Fig. 16 shown in [3]: starting with an image of a panda from ImageNet [4], the attacker adds a imperceptibly perturbation (i.e., an small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input), to make the image be recognised as a gibbon with high confidence by a particular deep neural net (GoogLeNet [3]). Also, recent research has shown that even in physical world scenarios, machine learning systems are vulnerable to adversarial examples: [4] shows how printed adversarial images (with modifications imperceptible to the human eye) obtained from a cell-phone camera are not correctly classified by the models. In general, adversarial examples have the potential to be dangerous. For example, attackers could target autonomous vehicles by using stickers or paint to create an adversarial stop sign that the vehicle would interpret as a ‘yield’ or other sign, as discussed in [5].

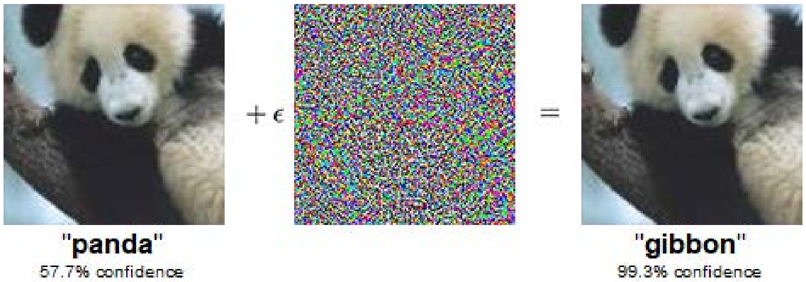


Fig. 16 An adversarial input, overlaid on a typical image, can cause a classifier to miscategorise a panda as a gibbon. Adapted from [3].

In response to concerns about the threats posed to a safe and trusted environment for AI technologies by adversarial attacks a field called adversarial machine learning has emerged over the past several years. Work in this area focuses on securing systems from disruptive perturbations at all points of vulnerability across the AI pipeline. One of the major safety strategies that has arisen from this research is an approach called model hardening, which has advanced techniques that combat adversarial attacks by strengthening the architectural components of the systems. Model hardening techniques may include adversarial training, where training data is methodically enlarged to include adversarial examples. Other model hardening methods involve architectural modification, regularisation, and data pre-processing manipulation. A second notable safety strategy is runtime detection, where the system is augmented with a discovery apparatus that can identify and trace in real-time the existence of adversarial examples. A valuable collection of resources to combat adversarial attack can be found at this link.

Bibliography

[1](1,2) Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[[2]] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[[3]] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9. 2015.

[[4]] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[[5]] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519. 2017.

This entry was readapted from *Leslie David. Understanding artificial intelligence ethics and safety. The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529, 2019* by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

Data Poisoning

In brief

Data poisoning occurs when an adversary modifies or manipulates part of the dataset upon which a model will be trained, validated, or tested. By altering a selected subset of training inputs, a poisoning attack can induce a trained AI system into curated misclassification, systemic malfunction, and poor performance. An especially concerning dimension of targeted data poisoning is that an adversary may introduce a ‘backdoor’ into the infected model whereby the trained system functions normally until it processes maliciously selected inputs that trigger error or failure. Data poisoning is possible because data collection and procurement often involves potentially unreliable or questionable sources. When data originates in uncontrollable environments like the internet, social media, or the Internet of Things, many opportunities present themselves to ill-intentioned attackers, who aim to manipulate training examples. Likewise, in third-party data curation processes (such as ‘crowdsourced’ labelling, annotation, and content identification), attackers may simply handcraft malicious inputs. [1]

More in detail

Data poisoning is a security threat to AI systems in which an attacker controls the behaviour of a system by manipulating its training, validation or testing data [4]. While it usually refers to the training data for machine learning algorithms, it could also affect some other AI systems by corrupting the testing data. Note that when the deployment data is corrupted during operation, we are in the situation of an Adversarial Attack. *Data_poisoning* is related to *data contamination*, although contamination is usually more accidental than intentional. For instance, many language models [8, 5, 6, 7]. are trained with data that is then used for test or validation, leading to an overoptimistic Evaluation of a system’s behaviour.

In the particular case of an attacker manipulating the training data by inserting incorrect or misleading information, as the algorithm learns from this corrupted data, it will draw unintended and even harmful conclusions. This type of threat is particularly relevant for deep learning systems because they require large amounts of data to train which is usually extracted from the web, and, at this scale, it is often infeasible to properly vet content. We find examples such as Imagenet [4] or the Open Images Dataset [8] containing tens or hundreds of millions of images from a wide range of potentially insecure and, in many cases, unknown sources. The current reliance of AI systems on such massive datasets that are not manually inspected has led to fears that corrupted training data can produce flawed models [9].

According to the breadth of the attack, data poisoning attacks fall into two main categories: attacks targeting *availability* and attacks targeting *integrity*. Availability attacks are usually unsophisticated but extensive, injecting as much erroneous data as possible into a database, so that the machine learning algorithm trained with this data will be totally inaccurate. Attacks against the integrity of machine learning are more complex and potentially more damaging. They leave most of the database intact, except for an imperceptible backdoor that allows attackers to control it. As a result, the model will apparently work as intended but with a fatal flaw. For instance, in a cybersecurity application, a classifier could make right predictions except when reading a specific file type, which is considered benign because hundreds of examples were included with that labelled in the corrupted dataset.

Depending on the timing of the attack, poisoning attacks can also be classified into two broad categories: *backdoor* and *triggerless poisoning attack*. The former causes a model to misclassify samples at test time that contain a particular trigger (e.g., small patches in images or characters sequence in text) [10, 11, 12, 13]. For example, training images could be manipulated so that a vision system does not identify any person wearing a piece of clothing having the trigger symbol printed on it. In this case model, the attacker modifies both the training data (placing poisons) and test data (inserting the trigger) [14, 15, 16]. Backdoor attacks cause a victim to misclassify any image containing the trigger. On the other hand, triggerless poisoning attacks do not require modifications at the time of inference and cause a victim to misclassify an individual sample [17].

Data poisoning attacks can cause considerable damage with minimal effort. Their effectiveness is almost directly proportional to the quality of the data. Poor quality data will produce subpar results, no matter how advanced the model is. For instance, the experiment ImageNet Roulette [18] used user-uploaded and labelled images to learn how to classify new images. Before long, the system began using racial and gender slurs to label people. Seemingly small and easily overlooked considerations, such as people using harmful language on the internet, become shockingly prevalent when an AI system learns from this data. As machine learning becomes more advanced, it will make more connections between data points that humans would not think of. As a result, even small changes to a database can have substantial repercussions.

While data poisoning is a concern, companies can defend against it with existing tools and techniques. The U.S. Department of Defense's Cyber Maturity Model Certification (CMMC) outlines four basic cyber principles for keeping machine learning data safe[2]: network (e.g., setting up and updating firewalls will help keep databases off-limits to internal and external threats), facility (e.g., restricting access to data centres), endpoint (e.g., use of data encryption, access controls and up-to-date anti-malware software) and people protection (e.g., user training). However, this assumes that the data is generated inside the limits of the organisation, but many training datasets are complemented with sources used for research or coming from social media, which are very difficult to vet. Also, with the current trend of using pretrained models and tuning them with smaller amounts of particular data, the risk is more on the data used for these pretrained models than unauthorised access to the finetuning data. Inspecting the models once trained, using techniques from explainable AI is also challenging, as the trapdoors may represent a very small percentage of the behaviour of the system. Overall, **data poisoning** is a complex problem that is closely related to other major problems in AI safety, and will remain problematic with the current paradigm of learning from massive amounts of data.

Bibliography

[[1]] Leslie David. Understanding artificial intelligence ethics and safety. *The Alan Turing Institute, https://doi.org/10.5281/zenodo.3240529*, 2019.

[[2]] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[[3]] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, 9389–9398. PMLR, 2021.

[[4]] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[[5]] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[[6]] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[[7]] Rishi Bommasani and others. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

[[8]] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and others. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[[9]] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, 2304–2313. PMLR, 2018.

[[10]] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[[11]] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019.

[[12]] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11957–11965. 2020.

[[13]] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.

[[14]] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

[[15]] W Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

[[16]] Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning*, 7614–7623. PMLR, 2019.

[[17]] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 2018.

[[18]] Kate Crawford and Trevor Paglen. Excavating ai: the politics of images in machine learning training sets. *AI and Society*, 2019.

This entry was written by Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares.

[[1]] Definition taken from [1] under Creative Commons Attribution License 4.0.

[[2]] https://cmmc-coe.org/test/

Fairness, Equity, and Justice by Design

In brief

The term fairness is defined as the quality or state of being fair; or a lack of favoritism towards one side. However, fairness can mean different concepts to different peoples, different contexts, and different disciplines [1]. An unfair Artificial Intelligence (AI) model produces results that are biased towards particular individuals or groups. The most relevant case of bias is discrimination against protected-by-law social groups. Equity requires that people are treated according to their needs, which does not mean all people are treated equally [2]. Justice is the “fair and equitable treatment of all individuals under the law” [1].

Abstract

We first provide motivations and background on fairness, equity and justice in AI. This consists of warnings and legal obligations about potential harms of unscrutinized AI tools, especially in socially sensitive decision making. A taxonomy of fair-AI algorithms is then presented, based on the step of the AI development process in which

fairness is checked/controlled for. Next, we summarize the guidelines and draft standards for fair-AI, and the software frameworks supporting the dimension. Finally, the main keywords of the dimension are extensively detailed.

Motivations and background [1]

Increasingly sophisticated algorithms from AI and Machine Learning (ML) support knowledge discovery from big data of human activity. They enable the extraction of patterns and profiles of human behavior which are able to make extremely accurate predictions. Decisions are then being partly or fully delegated to such algorithms for a wide range of socially sensitive tasks: personnel selection and wages, credit scoring, criminal justice, assisted diagnosis in medicine, personalization in schooling, sentiment analysis in texts and images, people monitoring through facial recognition, news recommendation, community bulding in social networks, dynamic pricing of services and products.

The benefits of algorithmic-based decision making cannot be neglected, e.g., procedural regularity – same procedure applied to each data subject. However, automated decisions based on profiling or social sorting may be biased [4] for several reasons. Historical data may contain human (cognitive) bias and discriminatory practices that are endemic, to which the algorithms assign the status of general rules. Also, the usage of AI/ML models reinforces such practices because data about model's decisions become inputs in subsequent model construction (feedback loops). Algorithms may wrongly interpret spurious correlations in data as causation, making predictions based on ungrounded reasons. Moreover, algorithms pursue the optimization of quality metrics, such as accuracy of predictions, that favor precision over the majority of people against small groups. Finally, the technical process of designing and deploying algorithms is not yet mature and standardized. Rather, it is full of small and big decisions (sometimes, trial and error steps) that may hide bias, such as selecting non-representative data, performing overspecialization of the models, ignoring socio-technical impacts, or using models in deployment contexts they are not tested for. These risks are exacerbated by the fact that the AI/ML models are complex for human understanding, or not even intelligible, sometimes they are based on randomness or time-dependent non-reproducible conditions [5].

Legal restrictions on automated decision-making are provided by the EU General Data Protection Regulation, which states (Article 22) "the right not to be subject to a decision based solely on automated processing". Moreover, (Recital 71) "in order to ensure fair and transparent processing in respect of the data subject […] the controller should use appropriate mathematical or statistical procedures […] to prevent, inter alia, discriminatory effects on natural persons".

Fair algorithms are designed with the purpose of preventing biased decisions in algorithmic decision making. Quantitative definitions have been introduced in philosophy, economics, and machine learning in the last 50 years [6, 7, 1], with more than 20 different definitions of fairness appeared thus far in the computer science literature [1, 10]. Four non-mutually exclusive strategies can be devised for fairness-by-design of AI/ML models.

Pre-processing approaches. They consists of a controlled sanitization of the data used to train an AI/ML model with respect to specific biases. Pre-processing approaches allow for obtaining less biased data, which can be used for training AI/ML models. An advantage of pre-processing approaches is that they are independent of the AI/ML model and algorithm at hand.

In-processing approaches. The second strategy is to modify the AI/ML algorithm, by incorporating fairness criteria in model construction, such as regularizing the optimization objective with a fairness measure. There is a fast growing adoption of in-processing approaches in many AI/ML problems other than in the original setting of classification, including ranking, clustering, community detection, influence maximization, distribution/allocation of goods, and models on non-structured data such as natural language texts and images. An area somehow in the middle between pre-processing and in-processing approaches is fair representation learning, where the model inferred from data is not used directly for decision making, but rather as intermediate knowledge.

Post-processing approaches. This strategy consists of post-processing an AI/ML model once it has been computed, so to identify and remove unfair decision paths. This can be achieved also by involving human experts in the exploration and interpretation of the model or of the model's decisions. Post-processing approaches consist of altering the model's internals, for instance by correcting the confidence of classification rules, or the probabilities of Bayesian models. Post-processing becomes necessary for tasks for which there is no in-processing approach explicitly designed for the fairness requirement at hand.

Prediction-time approaches. The last strategy assumes no change in the construction of AI/ML models, but rather correcting their predictions at run-time. Proposed approaches include promoting, demoting or rejecting predictions close to the decision boundary, differentiating the decision boundary itself over different social groups, or wrapping a fair classifier on top of a black-box base classifier. Such approaches may be applied to legacy software, including non-AI/ML algorithms, that cannot be replaced by in-processing approaches or changed by post-processing approaches.

Standards and guidelines

Several initiatives have started to audit, standardize and certify algorithmic fairness, such as the ICO Draft on AI Auditing Framework, the draft IEEE P7003™ Standard on Algorithmic Bias Considerations, the IEEE Ethics Certification Program for Autonomous and Intelligent Systems, and the ISO/IEC TR 24027:2021 Bias in AI systems and AI aided decision making (see also the entry on Auditing AI). Regarding the issue of equality data collection, the European Union High Level Group on Non-discrimination, Equality and Diversity has set up "Guidelines on improving the collection and use of equality data", and the European Union Agency for Fundamental Rights (FRA) maintains a list of promising practices for equality data collection.

Very few scientific works attempt at investigating the practical applicability of fairness in AI [11, 12]. This issue is challenging, and likely to require domain-specific approaches [13]. On the educational side, however, there are hundreds of university courses on technology ethics [14], many of which cover fairness in AI.

Software frameworks supporting dimension

The landscape of software libraries and tools is very large. Existing proposals cover almost every step of the data-friven AI development process (data collection, data processing, model development, model deployment, model monitoring), every type of AI models (classification, regression, clustering, ranking, community detection, influence maximization, distribution/allocation of goods), and every type of data (tabular, text, images, videos). Reviews and critical discussions of gaps for a few fairness toolkits can be found in [1, 16].

Main keywords

- Fairness, Equity, and Justice by Design: **Auditing AI** aims to identify and address possible risks and impacts while ensuring robust and trustworthy Accountability.
- Bias: **Bias** refers to an inclination towards or against a particular individual, group, or sub-groups. AI models may inherit biases from training data or introduce new forms of bias.
- Discrimination & Equity: Forms of bias that count as discrimination against social groups or individuals should be avoided, both from legal and ethical perspectives. Discrimination can be direct or indirect, intentional or unintentional.
- Fairness notions and metrics: The term **fairness** is defined as the quality or state of being fair; or a lack of favoritism towards one side. The notions of fairness, and quantitative measures of them (fairness metrics), can be distinguished based on the focus on individuals, groups and sub-groups.
- Fair Machine Learning: **Fair Machine Learning** models take into account the issues of bias and fairness. Approaches can be categorized as pre-processig, which transform the input data, as in-processing, which modify the learning algorithm, and post-processing, which alter models' internals or their decisions.
- Grounds of Discrimination: International and national laws prohibit **discriminating on some explicitly defined grounds**, such as race, sex, religion, etc. They can be considered in isolation, or interacting, giving rise to multiple discrimination and intersectional discrimination.
- Justice: **Justice** encompasses three different perspectives: (1) *fairness* understood as the fair treatment of people, (2) *rightness* as the quality of being fair or reasonable, and (3) a legal system, the scheme or system of law. Justice can be distinguished between *substantive* and *procedural*.
- Segregation: **Social segregation** refers to the separation of groups on the grounds of personal or cultural traits. Separation can be physical (e.g., in schools or neighborhoods) or virtual (e.g., in social networks).

Bibliography

[[1]] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. This thing called fairness: disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36, 2019.

[[2]] Martha Minow. Equality vs. Equity. *American Journal of Law and Equality*, 1:167–193, 2021.

[[3]] Jeffrey Lehman, Shirelle Phelps, and others. *West's encyclopedia of American law*. Thomson/Gale, 2004.

[[4]] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernández, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Bias in data-driven artificial intelligence systems - an introductory survey. *WIREs Data Mining Knowl. Discov.*, 2020.

[[5]] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *U. of Penn. Law Review*, 165:633–705, 2017.

[[6]] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: lessons for machine learning. In *FAT*, 49–58. ACM, 2019.

[[7]] Reuben Binns. Fairness in machine learning: lessons from political philosophy. In *FAT*, volume 81 of Proceedings of Machine Learning Research, 149–159. PMLR, 2018.

[[8]] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, 29(5):582–638, 2014.

[[9]] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2021.

[[10]] Indre Zliobaite. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.*, 31(4):1060–1089, 2017.

[[11]] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *SIGKDD Explor.*, 23(1):14–23, 2021.

[[12]] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. Putting fairness principles into practice: challenges, metrics, and improvements. In *AIES*, 453–459. ACM, 2019.

[[13]] Michelle Seng Ah Lee and Luciano Floridi. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds Mach.*, 31(1):165–191, 2021.

[[14]] Casey Fiesler, Natalie Garrett, and Nathan Beard. What do we teach when we teach tech ethics?: A syllabi analysis. In *SIGCSE*, 289–295. ACM, 2020.

[[15]] Michelle Seng Ah Lee and Jatinder Singh. The landscape and gaps in open source fairness toolkits. In *CHI*, 699:1–699:13. ACM, 2021.

[[16]] Brianna Richardson and Juan E. Gilbert. A framework for fairness: A systematic review of existing fair AI solutions. *CoRR*, 2021.

[[17]] Salvatore Ruggieri. Algorithmic fairness. In *Elgar Encyclopedia of Law and Data Science*. Edward Elgar Publishing Limited, 2022.

This entry was written by Salvatore Ruggieri.

[1] This Section was readapted from [17].

Auditing AI

In brief

Auditing AI aims to identify and address possible risks and impacts while ensuring robust and trustworthy Accountability.

More in Detail

One of the measures to ensure that AI is used responsibly is the initiation of auditing practices as they facilitate to verify if the system works as intended.

Audits can be conducted either in-house or by external parties. The former requires internal evaluation regarding whether systems are fit, the human elements involved with the system are appropriate and monitored, and the technical elements of the system are in perfect condition and function correctly [2, 3]. The latter involves both regulators and third-parties verifying compliance [3]. Interestingly, critical external audits encompasses “disparate methods of journalist, technicians, and social scientist who have examined the consequences of already-deployed algorithmic systems and who have no formal relationship which the institutions designing or integrating the audited systems” [5]. Well-known examples of these practices such as the Propublica’s examination of Northpoint recivism prediction API [6] or the Gender Shade Project [7], have played a crucial role pointing out harmful application of algorithm systems to draw the attention of the society and require companies an active role setting out governance and accountability mechanism.

As a result of these social demands, internal governance mechanisms [5] have been introduced from within the own companies that design and deployed the algorithmic systems. The goal is to propose technical and organisational procedures, among which are detailed frameworks for algorithm auditing [8], able to identify and address possible risk and impacts while ensuring robust and trustful accountability. In essence, precise and well-documented audits facilitate later scrutiny offering records on the reasons for the audit to be initiated, the procedures that were followed as well as the conclusions that were reached and, if carried out, the remedies or measures that were adopted.

To this regard, more and more voices consider audits as indispensable accountability mechanism to ensure the compliance of AI systems along their life-cycle with the different applicable legislation, concerning in particular privacy and data protection law [9]. Moreover, AI auditing can benefit from extensive literature in more mature disciplines, such as audit studies in social sciences [10] and empirical economics [1]. Audits facilitate private entities the provision of documentation when requested by public bodies, favouring a systematic governance [11] of AI systems through a general transparency and enforcement regime. This joint effort between public and private institutions would, in turn, result in collaborative governance scheme [11].

The upcoming EU Artificial Intelligence Act can be seen as a proposal to establish a Europe-wide ecosystem for conducting AI auditing [12] and in line with that idea more and more research is done on auditing procedures for algorithms (for reviews see [13, 14]). For example, [8] propose a framework for internal AI auditing which includes both ethical aspects (a social impact assessment and ethical risk analysis chart) and technical audits (such as adversarial testing and a Failure Modes and Effect Analysis). Such audits are often supported by technical documentation, such as the Datasheets for Datasets proposal [15] to maintain information on datasets used to train AI systems. Such documentation can both help to ensure that AI systems are deployed for tasks in line with the data they were trained on and help to spot ethical risks stemming from the data [16], such as biases.

Ethical risks can also be the sole focus of AI audits, as in ethics-based auditing (proposed for AI in [17]). While still in development, several options are emerging where: “functionality audits focus on the rationale behind decisions; code audits entail reviewing the source code of an algorithm; and impact audits investigate the types, severity, and prevalence of effects of an algorithm’s outputs.” [18] For these audits in particular determining what is measured can be a challenge, as it is difficult to define clear metrics on which ethical aspects of AI systems can be evaluated. *Fairness metrics* (cf. the entry on) can certainly help here, but as discussed there is a difficulty in the selection of the right metric and even then there are limitations and trade-offs with other metrics. In addition, for the integration of AI ethics in ESG (Environmental, Social and Governance) reporting towards investors [19] such fairness metrics need not give sufficient insights into whether algorithms are used responsibly at an organisational level. Existing ESG criteria for organizational audits may help here, as well as work on KPIs for Responsible Research and Innovation [20]. Despite all this work on metrics, it is however still an open question to what extent ethics can be captured in numbers the way other aspects of audits are, with some arguing that it is impossible to develop benchmarks for how ethical an AI system is [21]. Instead, they argue, the focus should be on values and value trade-offs.

Z-Inspection, another auditing framework proposed based on the European High Level Expert Group's Guidelines for Trustworthy AI [22], takes values as its starting point [23]. As can also be seen in a case study for the framework involving an algorithm that recognizes cardiac arrests in emergency calls [24] this framework proceeds from a wide identification of stakeholders and their values to the analysis of (socio-)technical scenario's to reach an identification and (potentially) resolution of ethical, technical and legal issues of an AI system. Ultimately this still depends on the translation of values into metrics, and so the main challenge of developing such metrics stands regardless of one's auditing approach.

Standards represent a natural framework for the proceduralization of audits. Certification by neutral third party states compliance to certain standards as the result of auditing. Several draft proposals are being prepared which include (at least implicitely) elements for conducting audits, such as the following:

- ISO/IEC TR 24027:2021 - Artificial intelligence (AI) — Bias in AI systems and AI aided decision making.
- IEEE 7010-2020: Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being.
- IEEE P2863 - Recommended Practice for Organizational Governance of Artificial Intelligence.
- IEEE CertifAIEd – Ontological Specification for Ethical Accountability
- IEEE CertifAIEd – Ontological Specification for Ethical Algorithmic Bias.
- NIST AI Risk Management Framework.

However, there is not yet a formal professional standard to guide auditors of AI systems, yet some guidelines exist.

Bibliography

[[1]] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, 29(5):582–638, 2014.

[[2]] Ada Lovelace and UK DataKind. Examining the black box: tools for assessing algorithmic systems. Technical Report, Technical report, AdaLovelace Institute, 2020. URL: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.

[[3]] Emre Kazim, Danielle Mendes Thame Denny, and Adriano Koshiyama. Ai auditing and impact assessment: according to the uk information commissioner's office. *AI and Ethics*, 1(3):301–310, 2021.

[[4]] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Reviewable automated decision-making: a framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. 2021.

[5](1,2) Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic impact assessments and accountability: the co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746. 2021.

[[6]] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

[[7]] Joy Adowaa Buolamwini. *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017.

[8](1,2)

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *FAT**, 33–44. ACM, 2020.

[[9]] Bryan Casey, Ashkon Farhangi, and Roland Vogl. Rethinking explainable machines: the gdpr's' right to explanation'debate and the rise of algorithmic audits in enterprise. *Berkeley Tech. LJ*, 34:143, 2019.

[[10]] Briana Vecchione, Karen Levy, and Solon Barocas. Algorithmic auditing and social justice: lessons from the history of audit studies. In *EAAMO*, 19:1–19:9. ACM, 2021.

[11](1,2) Margot E Kaminski and Gianclaudio Malgieri. Algorithmic impact assessments under the GDPR: producing multi-layered explanations. *International Data Privacy Law*, pages 19–28, 2020.

[[12]] Jakob Mökander, Maria Axente, Federico Casolari, and Luciano Floridi. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed european ai regulation. *Minds and Machines*, pages 1–28, 2021.

[[13]] Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, and others. Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. *SSRN Electronic Journal*, 2021.

[[14]] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, and others. Auditing algorithms: understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344, 2021.

[[15]] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[[16]] Karen L Boyd. Datasheets for datasets help ml engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[[17]] Jakob Mökander and Luciano Floridi. Ethics-based auditing to develop trustworthy ai. *Minds and Machines*, 31(2):323–327, 2021.

[[18]] Jakob Mökander, Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. Ethics-based auditing of automated decision-making systems: nature, scope, and limitations. *Science and engineering ethics*, 27(4):1–30, 2021.

[[19]] Matti Minkkinen, Anniina Niukkanen, and Matti Mäntymäki. What about investors? ESG analyses as tools for ethics-based AI auditing. *AI & SOCIETY*, pages 1–15, 2022.

[[20]] Zenlin Kwee, Emad Yaghmaei, and Steven Flipse. Responsible research and innovation in practice an exploratory assessment of key performance indicators (kpis) in a nanomedicine project. *Journal of Responsible Technology*, 5:100008, 2021.

[[21]] Travis LaCroix and Alexandra Sasha Luccioni. Metaethical perspectives on'benchmarking'ai ethics. *arXiv preprint arXiv:2204.05151*, 2022. URL: https://arxiv.org/abs/2204.05151.

[[22]] Nathalie Smuha. The EU approach to ethics guidelines for trustworthy Artificial Intelligence. In *Computer Law Review International*, volume 20, 97–106. 2019.

[[23]] Roberto V Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, and others. Z-inspection®: a process to assess trustworthy AI. *IEEE Transactions on Technology and Society*, 2(2):83–97, 2021.

[[24]] Roberto V Zicari, James Brusseau, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Marianna B Ganapini, Sara Gerke, Thomas Krendl Gilbert, Eleanore Hickman, Elisabeth Hildt, and others. On assessing trustworthy ai in healthcare. machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Frontiers in Human Dynamics*, pages 30, 2021.

This entry was written by Alejandra Bringas Colmenarejo, Stefan Buijsman, and Salvatore Ruggieri.

Bias

In brief

Bias refers to an inclination towards or against a particular individual, group, or sub-groups. AI models may inherit biases from training data or introduce new forms of bias.

More in Detail

The success of Machine Learning (ML) systems in visual recognition, online advertising, and recommendation systems have inspired its use in applications such as employee hiring, legal systems, social systems, and voice interfaces such as Alexa, Siri, and the like. Along with the proliferation of these domains, a significant concern regarding the trustworthiness of decisions has risen due to various biases (or systematic errors) which may produce skewed results in the automated decision making. The word ‘bias’ has an established normative explanation in legal language, where it refers to ‘judgement based on preconceived notions or prejudices, as opposed to the impartial evaluation of facts’ [2]. In a more generalized version, bias refers to an inclination towards or against a particular individual, group, or sub-groups. The real world is often described as biased in this sense, and since machine learning techniques simply imitate observations of the world, it should come as no surprise that the resulting systems also capture the same bias [3].

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) [4] software for recidivism prediction, used by the U.S department courts to decide whether to release a person or keep them in prison, has discontinued its use after a careful investigation conducted by the U.S officers as they concluded that the software is biased against African-Americans. Also, the use of predictive policing [5] software has been ceased due to the presence of racial biases. Amazon’s employment hiring [6] application realized that it is biased against women. Content personalization and ad ranking systems have been accused of filter bubbles and racial and gender profiling. Bidirectional Encoder Representations from Transformers (BERT), has shown signs of gender biases in google search as it is observed that the gender-neutral terms (such as receptionist, doctor, nurse etc.) acquire stereotype and bias due to the context in which they are present in the corpus [7]. In image domain, a latest gender classification report from the National Institute for Standards and Technology (NIST) pointed out that image classification algorithms performed worse for female-labeled faces than male-labeled faces , exhibit gender biases [8]. A bias can exist in different forms and shapes based on the domain and context of application. The main reasons for the origin of these biases are manifold. An outline of bias-inducing stages in the ML pipeline is detailed in [1]. Based on this study, bias definitions can be induced in data, algorithms, and user interaction feedback loops.

ML systems are primarily based on data-driven approaches; therefore, the outcome of ML-based decision-making processes depends on the input data and the interpretation of that data. This decision-making process involves numerous data analyses, such as uncovering patterns in the data, finding correlations and trends, missing data imputations, and data pre-processing. The performance of ML models depends on the data used to train these models and the analysis performed on the training data. It is noted that the primary source of biases is from the data and its processing- involves what data was used for training, how it was collected, and how the data was generated and pre-processed. A general definition of dataset bias is that the data is not representative of the population of study [9]. Nevertheless, in a broader sense, it also occurs when the data does not contain features for specific applications we are interested in. Additionally, human interactions with the data produce bias against a specific group or individual [9]. Various forms of dataset biases have been identified in ML systems. Sample/selection biases emerge due to the non-random sampling of groups and sub-groups. Exclusion bias arises at the data pre-processing stage when valuable data are omitted thought to be unnecessary. When the data used for training a model is different from the data collected from the real world, for example, the training data is collected using a fixed camera in image training, but the production data is collected using different cameras, a measurement bias can be occurred. Recall bias is a kind of measurement bias, and it occurs when similar data are inconsistently labeled. Observer bias occurs when we observe data based on what we want to see or expect to see. Association biases are resultant of the spurious correlations between features in the data.

Furthermore, algorithmic biases are systematic errors in computer systems or models that cause certain privileges in outcomes concerning a particular group or a person. These biases can emerge in various ways. Foremost among these are the design of the algorithm or the way it uses the datasets to be coded, collected, selected, and processed. Algorithmic errors may lead to biased outcomes even though the data used for training are unbiased. A clear example is pre-existing bias, arising as the result of underlying social and institutional ideologies [10]. Another algorithmic bias is caused by technical biases manifested due the technical limitations of code, its computational resources, its design, and the constraints on the system. Technical biases are more frequent when we rely more on the algorithm in other domains or unanticipated contexts. Moreover, and related to the algorithm internals, correlation biases materialise when algorithms assume conclusions from the correlations in data attributes without knowing the specific purpose of those attributes. Finally, another known bias – in between the identified ones – are feedback loop biases. These arise when there is a recursion error in the mechanism in which information is processed into the data-model-experience pipeline.

The skewed outcomes from the biased data or (and) biased algorithms affect user decisions which may result in a more biased data for future ML systems. For example, consider a search engine which ranks queries. The end users interact mostly with the top ranked results, rather than going down the list, that can affect popularity and user interest of the upcoming decisions, due to the biased interactions.

As a long term vision to create responsible ML systems, identifying and mitigating biases throughout the ML development life cycle should be given paramount importance. In a broader sense, the different ways the bias could be mitigated are:

1. Identify and define potential sources
2. Set up guidelines and rules for data collection as well as a model use
3. Define accurate representative data for training
4. Properly document and share how the entire data collection process has been done
5. Incorporate ways to measure and mitigate biases as part of the standard evaluation procedure.

Bibliography

[[1]] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2021.

[[2]] A Campolo, M Sanfilippo, M Whittaker, and K Crawford. Ai now 2017 symposium and workshop. *AI Now Institute at New York University*, 2018.

[[3]] Thomas Hellström, Virginia Dignum, and Suna Bensch. Bias in machine learning–what is it good for? *arXiv preprint arXiv:2004.00686*, 2020. URL: https://arxiv.org/abs/2004.00686.

[[4]] Tim Brennan and William Dieterich. Correctional offender management profiles for alternative sanctions (compas). *Handbook of Recidivism Risk/Needs Assessment Tools (2018)*, 2018.

[[5]] Kiana Alikhademi, Emma Drobina, Diandra Prioleau, Brianna Richardson, Duncan Purves, and Juan E Gilbert. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, pages 1–17, 2021.

[[6]] Akhil Alfons Kodiyan. An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool. *Researchgate Preprint*, 2019.

[[7]] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: measuring stereotypical bias in pretrained language models. In *ACL/IJCNLP (1)*, 5356–5371. Association for Computational Linguistics, 2021.

[[8]] Patrick J Grother, Patrick J Grother, and Mei Ngan. *Face recognition vendor test (frvt)*. US Department of Commerce, National Institute of Standards and Technology, 2014.

[9](1,2) Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[[10]]

Roel Dobbe, Sarah Dean, Thomas Krendl Gilbert, and Nitin Kohli. A broader view on bias in automated decision-making: reflecting on epistemology and dynamics. *CoRR*, 2018. arXiv:1807.00553.

This entry was written by Resmi Ramachandranpillai, Fredrik Heintz, Miguel Couceiro, and Gabriel Gonzalez-Castañé.

Discrimination & Equity

In brief

Forms of bias that count as discrimination against social groups or individuals should be avoided, both from legal and ethical perspectives. Discrimination can be direct or indirect, intentional or unintentional.

More in Detail

Not all forms of *bias* (also known as *statistical discrimination*) are problematic. Here we use the normative sense of *discrimination*, where all forms of bias that count as discrimination are considered problematic and should be avoided. This discrimination can be direct (where a protected feature is intentionally used in the decision making procedure). In such a case explainability tools such as feature importance methods can help to detect whether a model's decisions are based on the feature, in addition to the fairness metrics in the entry on . Often, however, protected features are purposely not included among the input variables, and so no direct discrimination will take place. Instead, indirect discrimination (where there is direct discrimination on features that strongly correlate with a protected feature, in such a way that users with a socially salient value of the feature – e.g. women – are worse off, cf. [1]) is the most common type of discrimination in machine learning systems. For ways to detect these cases of indirect discrimination, see the fairness metrics.

This type of indirect discrimination is often unintentional. Philosophical accounts thus disagree about the degree of intentionality that is required for bias to count as discrimination: mental state accounts [2] require systematic animosity or preferences towards a certain group. Such animosity need not be present among the designers of the system, though it may be part of the reason for the societal biases that filter through into the data. Other accounts [1, 2] opt for weaker notions of intentionality, where it is sufficient to enable the feature/group membership to play a role in the decision making procedure. This clearly allows for the (frequent) scenario where an algorithm has disparate impacts on groups even when this was not the result of preferences/animosities of the developers. Yet even then not all types of bias are considered normatively problematic: a statistical bias that negatively impacts smokers is not clearly a case of discrimination. So when is a bias a case of discrimination?

An influential point of departure is the idea that biases should not be on features outside of people's control [4, 5]. This might explain why the paradigmatic cases of discrimination is when there is disparate treatment based on gender, race, or ethnicity (cf. the entry Grounds of Discrimination), as we cannot choose these. However, there are more features beyond our control, as illustrated by the 'other people's choice principle' [4]: statistical patterns resulting from other people's choices are not in our control either, and thus may lead to discrimination. Consequently, charging higher premiums to a buyer of a red car because on average drivers of red cars cause more accidents may be seen as problematic. It violates the other people's choice principle, as a buyer has no control over the driving of other car owners. On the other hand, charging higher premiums to smokers does not violate this principle, since smoking is a direct cause of higher health risks. In practice, however, it is difficult to draw a clear border, as e.g. socio-economic status impacts people's choices. Instead, some authors [2] suggest to consider notions of Justice as guiding the distribution of benefits and burdens resulting from the use of AI. For example, luck egalitarianism would consider it discriminatory to uphold biases which reflect factors of luck. Still, while the exact confines of normatively problematic discrimination are difficult to define, it is clear that gender, race, etc. are protected features and that such discrimination needs to be detected and tackled.

Following the egalitarian principles of [7, 8], some authors address fairness from a multi-agent perspective [9, 10] in automated decision making. Taking a welfare perspective [11] propose a family of welfare based measures that can be integrated together with other fairness and performance constraints. Following the same tracks, [12] considered the temporal/sequential dimension and addressed fairness in the context of Markov decision processes and reinforcement learning. Such multi-objective and welfare approaches not only enforce human intervention and social criteria to prevent unfair outcomes for some users or stakeholders, but could be adapted to ensure equity and fair trade-off between privilege and unprivileged groups.

Bibliography

[1](1,2) Kasper Lippert-Rasmussen. *Born free and equal?: A philosophical inquiry into the nature of discrimination*. Oxford University Press, 2013.

[[2]] Thomas M Scanlon. Moral dimensions. In *Moral Dimensions*. Harvard University Press, 2009.

[[3]] Michele Loi and Markus Christen. Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, 34(4):967–992, 2021.

[4](1,2) Kasper Lippert-Rasmussen. Nothing personal: on statistical discrimination. *Journal of Political Philosophy*, 15(4):385–403, 2007.

[[5]] Daniel E Palmer. Insurance, risk assessment and fairness: an ethical analysis. In *Insurance ethics for a more ethical world*. Emerald Group Publishing Limited, 2007.

[[6]] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In *CHI*, 377. ACM, 2018.

[[7]] John Rawls. *A theory of justice*. Harvard university press, 2009.

[[8]] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.

[[9]] Steven de Jong, Karl Tuyls, and Katja Verbeeck. Fairness in multi-agent systems. *Knowl. Eng. Rev.*, 23(2):153–180, 2008.

[[10]] Jianye Hao and Ho-fung Leung. *Fairness in Cooperative Multiagent Systems*, pages 27–70. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[[11]] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *NeurIPS*, 1273–1283. 2018.

[[12]] Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *ICML*, volume 70 of Proceedings of Machine Learning Research, 1617–1626. PMLR, 2017.

This entry was written by Stefan Buijsman and Miguel Couceiro.

Fairness notions and metrics

In brief

The term **fairness** is defined as the quality or state of being fair; or a lack of favoritism towards one side. The notions of fairness, and quantitative measures of them (fairness metrics), can be distinguished based on the focus on individuals, groups and sub-groups.

More in Detail

The term fairness is defined as the quality or state of being fair; or a lack of favoritism towards one side. However, like Bias, fairness can mean different concepts to different peoples, different contexts, and different disciplines. The definition of fairness in various disciplines is detailed in [1]. An unfair model produces results that are skewed towards particular individuals or groups. The primary sources of this unfairness are the presence of biases. There are two important categories of biases which play crucial role in fairness; (i) technical bias and (ii) social bias. Technical biases can be traced back to the sources , but social biases are very difficult to fix as these are a matter of politics, perspectives, and shifts in prejudices and preconceptions that can take years to change [3]. Most of the state-of-the-art techniques tackle technical errors, but it cannot resolve the root causes of bias. Based on this observation, Sandra et al. [3] proposed three responses concerning algorithmic bias and resulting social inequality. The first is not an active choice as it allows the system to get worse and do nothing to fix biases. Second, incorporate techniques to fix technical errors and maintain a status quo to ensure that the system do not make it worse. Much works in fairness focused on this option, called ‘bias preserving fairness’, maintains a status quo as a baseline, aligns with the formal equality of EU non-discrimination law. Finally, ‘bias transforming fairness’, the third response focuses on the substantive equality of EU non-discrimination which can only be achieved by accounting for historical (social) inequalities. As argued in [3], users (developers,deployers etc.) should give preference to ‘bias transforming’ fairness metrics, when a fairness metric is used to make substantive decisions about people in contexts where significant disparity has been previously observed.

The notions of fairness fall under individuals, groups and sub-groups. Individual fairness ensures that similar individuals should be treated similarly. It accounts for the distance measures to evaluate the similarity of individuals [4, 5]. On the other hand, group fairness compares quantities at the group level primarily identified by protective features such as gender, ethnicity etc. etc. [6, 7]. Sub-group fairness is more rigid than group fairness as this ensures fairness concerning one or more structured sub-groups defined by sensitive features, interpolates between individual and group fairness notions [8]. According to [9], it is impossible to satisfy all of the above notions, leading to conflicts between fairness definitions. Therefore, one suggestion could be to select appropriate fairness criteria and use those based on the application and deployment. Another concern has risen in [10], temporal aspects of fairness notions may harm the sensitive groups over time if not updated.

Some widely used fairness metrics: In order to recall some widely used fairness metrics we need to introduce some notation. Let \(V\), \(A\), and \(X\) be three random variables representing, respectively, the total set of features, the sensitive features, and the remaining features describing an individual such that \(V=(X,A)\) and \(P(V=v_i)\) represents the probability of drawing an individual with a vector of values \(v_i\) from the population. For simplicity, we focus on the case where \(A\) is a binary random variable where \(A=0\) designates the protected group, while \(A=1\) designates the non-protected group. Let \(Y\) represent the actual outcome and \(\hat{Y}\) represent the outcome returned by the prediction algorithm. Without loss of generality, assume that \(Y\) and \(\hat{Y}\) are binary random variables where \(Y=1\) designates a positive instance, while \(Y=0\) a negative one. Typically, the predicted outcome \(\hat{Y}\) is derived from a score represented by a random variable \(S\) where \(P[S = s]\) is the probability that the score value is equal to \(s\).

Statistical parity [11] is one of the most commonly accepted notions of fairness. It requires the prediction to be statistically independent of the sensitive feature \((\hat{Y} \perp A)\). In other words, the predicted acceptance rates for both protected and unprotected groups should be equal. Statistical parity implies that
\(\displaystyle \frac{TP+FP}{TP+FP+FN+TN}\) [1]
is equal for both groups. A classifier Ŷ satisfies statistical parity if:
\(\label{eq:sp} P[\hat{Y} \mid A = 0] = P[\hat{Y} \mid A = 1].\)

Conditional statistical parity [12] is a variant of statistical parity obtained by controlling on a set of resolving features[2]. The resolving features (we refer to them as \(R\)) among \(X\) are correlated with the sensitive feature \(A\) and give some factual information about the label at the same time leading to a *legitimate* discrimination. Conditional statistical parity holds if:
\(\label{eq:csp} P[\hat{Y}=1 \mid R=r,A = 0] = P[\hat{Y}=1 \mid R=r,A = 1] \quad \forall r \in range(R).\)

Equalized odds [13] considers both the predicted and the actual outcomes. The prediction is conditionally independent from the protected feature, given the actual outcome \((\hat{Y} \perp A \mid Y)\). In other words, equalized odds requires that both sub-populations to have the same true positive rate \(TPR = \frac{TP}{TP+FN}\) and false positive rate \(FPR = \frac{FP}{FP+TN}\):
\(\label{eq:eqOdds} P[\hat{Y} = 1 \mid Y=y,\; A=0] = P[\hat{Y}=1 \mid Y= y,\; A=1] \quad \forall{ y \in \{0,1\}}.\)

Because equalized odds requirement is rarely satisfied in practice, two variants can be obtained by relaxing its equation. The first one is called **equal opportunity** [13] and is obtained by requiring only TPR equality among groups:

\(\label{eq:eqOpp} P[\hat{Y}=1 \mid Y=1,A = 0] = P[\hat{Y}=1\mid Y=1,A = 1].\)
As \(TPR\) does not take into consideration \(FP\), equal opportunity is completely insensitive to the number of false positives.

The second relaxed variant of equalized odds is called **predictive equality** [12] which requires only the FPR to be equal in both groups:
\(\label{eq:predEq} P[\hat{Y}=1 \mid Y=0,A = 0] = P[\hat{Y}=1\mid Y=0,A = 1].\)
Since \(FPR\) is independent from \(FN\), predictive equality is completely insensitive to false negatives.

Conditional use accuracy equality [14] is achieved when all population groups have equal positive predictive value \(PPV=\frac{TP}{TP+FP}\) and negative predictive value \(NPV=\frac{TN}{FN+TN}\). In other words, the probability of subjects with positive predictive value to truly belong to the positive class and the probability of subjects with negative predictive value to truly belong to the negative class should be the same. By contrast to equalized odds, one is conditioning on the algorithm's predicted outcome not the actual outcome. In other words, the emphasis is on the precision of prediction rather than its recall:
\(\label{eq:condUseAcc} P[Y=y\mid \hat{Y}=y ,A = 0] = P[Y=y\mid \hat{Y}=y,A = 1] \quad \forall{ y \in \{0,1\}}.\)

Predictive parity [15] is a relaxation of conditional use accuracy equality requiring only equal \(PPV\) among groups: $\(\label{eq:predPar} P[Y=1 \mid \hat{Y} =1,A = 0] = P[Y=1\mid \hat{Y} =1,A = 1]\)$ Like predictive equality, predictive parity is insensitive to false negatives.

Overall accuracy equality [14] is achieved when overall accuracy for both groups is the same. This implies that
\[\label{eq:accuracy} \frac{TP+TN}{TP+FN+FP+TN}\]
is equal for both groups:
\[\label{eq:ovAcc} P[\hat{Y} = Y | A = 0] = P[\hat{Y} = Y | A = 1]\]
Treatment equality [14] is achieved when the ratio of FPs and FNs is the same for both protected and unprotected groups:
\(\label{eq:treatEq} \frac{FN}{FP}\)$_{A=0}$ \(= \frac {FN}{FP}\)$_{A=1}$

Total fairness [14] holds when all aforementioned fairness notions are satisfied simultaneously, that is, statistical parity, equalized odds, conditional use accuracy equality (hence, overall accuracy equality), and treatment equality. Total fairness is a very strong notion which is very difficult to hold in practice.

Balance [9] uses the score (\(S\)) from which the outcome \(Y\) is typically derived through thresholding.
Balance for positive class focuses on the applicants who constitute positive instances and is satisfied if the average score \(S\) received by those applicants is the same for both groups:
\(\label{eq:balPosclass} E[S \mid Y =1,A = 0)] = E[S \mid Y =1,A = 1].\)
Balance of negative class focuses instead on the negative class:
\(\label{eq:balNegclass} E[S \mid Y =0,A = 0] = E[S \mid Y =0,A = 1].\)

Calibration [15] holds if, for each predicted probability score \(S=s\), individuals in all groups have the same probability to actually belong to the positive class:
\(\label{eq:calib} P[Y =1 \mid S =s,A = 0] = P[Y =1 \mid S =s,A = 1] \quad \forall s \in [0,1].\)

Well-calibration [9] is a stronger variant of calibration. It requires that (1) calibration is satisfied, (2) the score is interpreted as the probability to truly belong to the positive class, and (3) for each score \(S=s\), the probability to truly belong to the positive class is equal to that particular score:
\(\label{eq:wellCalib} P[Y =1 \mid S =s,A = 0] = P[Y =1 \mid S =s,A = 1] = s \quad \forall \; {s \in [0,1]}.\)

Fairness through awareness [11] implies that similar individuals should have similar predictions. Let \(i\) and \(j\) be two individuals represented by their attributes values vectors \(v_i\) and \(v_j\). Let \(d(v_i,v_j)\) represent the similarity distance between individuals \(i\) and \(j\). Let \(M(v_i)\) represent the probability distribution over the outcomes of the prediction. For example, if the outcome is binary (\(0\) or \(1\)), \(M(v_i)\) might be \([0.2,0.8]\) which means that for individual \(i\), \(P[\hat{Y}=0]) = 0.2\) and \(P[\hat{Y}=1] = 0.8\). Let \(d_M\) be a distance metric between probability distributions. Fairness through awareness is achieved iff, for any pair of individuals \(i\) and \(j\):
\(d_M(M(v_i), M(v_j)) \leq d(v_i, v_j)\)
In practice, fairness through awareness assumes that the similarity metric is known for each pair of

individuals [16]. That is, a challenging aspect of this approach is the difficulty to determine what is an appropriate metric function to measure the similarity between two individuals. Typically, this requires careful human intervention from professionals with domain expertise [17].

Process fairness [18] (or procedural fairness) can be described as a set of subjective fairness notions that are centered on the process that leads to outcomes. These notions are not focused on the fairness of the outcomes, instead they quantify the fraction of users that consider fair the use of a particular set of features. They are subjective as they depend on user judgments which may be obtained by subjective reasoning.

A natural approach to improve process fairness is to remove all sensitive (protected or salient) features before training classifiers. This simple approach connects process fairness to *fairness through unawareness*. However, there is a trade-off to manage since dropping out sensitive features may impact negatively classification performance [19].

Nonstatistical fairness metrics: Recently, further metrics have been proposed and that differ from the previous in that they do not fully rely on statistical considerations, and take into account domain knowledge, that is not directly observable from data, require expert input, or reason about hypothetical situations. As they fall out of the scope of this chapter, we will not further dwell into these and simply mention a few to the interested reader: *total effect* [20] (that is the “causal” version of statistical parity and measures the effect of changing the value of an attribute, taking into account a given causal graph), *effect of treatment of the treated* [20] (that relies on counterfactuals with respect to sensitive features and measures the difference between the probabilities of instances and their counterfactuals), and *counterfactual fairness* [17] (which is a fine-grained variant of the previous but with respect to the set all features).

Discussion: As the above fairness metrics often conflict, and it is not possible to be fair according to all of these definitions, it is a challenge to choose the relevant metric to focus on. While still very much an open research area, some suggestions on how one can deal with conflicts between fairness metrics can be found in [2, 21]. Indeed, fairness metrics frequently conflict with other metrics such as accuracy and privacy. [22] show that in a credit scoring case enforcing fairness metrics can lead to significant drops in accuracy and, thus, maximum profit. This is unavoidable: improvements on fairness often result in lower accuracy, and research on the Pareto frontier for this trade-off is now emerging [23, 24]. Similarly, there is a trade-off between fairness and privacy, as fairness metrics typically require sensitive information in order to be used. As a result, fairness affects privacy (and vice versa), for example in facial recognition [25] and medical applications [26].

Finally, there is a connection between *fairness* and Justice, seen in for example Rawls’ work on Justice as Fairness [27]. And indeed, a range of theories of (distributive) justice describe how benefits and burdens should be distributed (cf. the entry on). As such, they can be seen as guiding the outcomes of algorithms even if they describe what these distributions should be in society as a whole. Yet, as [3] argue at length, there is little overlap between theories of distributive justice and fairness metrics. Non-comparative notions of justice are not captured by fairness metrics, nor are notions such as Rawls’ difference principle, on which the right distribution is the one where the worst off have the highest absolute level of benefits. Fairness metrics have focused more on discrimination than on *justice*.

Bibliography

[[1]] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. This thing called fairness: disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36, 2019.

[[2]] Michele Loi and Markus Christen. Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosophy & Technology*, 34(4):967–992, 2021.

[3](1,2,3) Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

[[4]] Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *UAI*, volume 124 of Proceedings of Machine Learning Research, 749–758. AUAI Press, 2020.

[[5]] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: amortizing individual fairness in rankings. In *SIGIR*, 405–414. ACM, 2018.

[[6]] Yu Cheng, Zhihao Jiang, Kamesh Munagala, and Kangning Wang. Group fairness in committee selection. *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–18, 2020.

[[7]] Vincent Conitzer, Rupert Freeman, Nisarg Shah, and Jennifer Wortman Vaughan. Group fairness for the allocation of indivisible goods. In *AAAI*, 1853–1860. AAAI Press, 2019.

[[8]] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In *ICML*, volume 80 of Proceedings of Machine Learning Research, 2569–2577. PMLR, 2018.

[9](1,2,3) Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, volume 67 of LIPIcs, 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[[10]] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *IJCAI*, 6196–6200. ijcai.org, 2019.

[11](1,2) Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. 2012.

[12](1,2) Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *KDD*, 797–806. ACM, 2017.

[13](1,2) Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, 3315–3323. 2016.

[14](1,2,3,4) Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods & Research*, 50:3–44, 2018.

[15](1,2) Alexandra Chouldechova. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[[16]] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. In *NeurIPS*, 4847–4857. 2018.

[17](1,2) Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, 4066–4076. 2017.

[[18]] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. In *AAAI*, 51–60. AAAI Press, 2018.

[[19]] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In *WWW*, 1171–1180. ACM, 2017.

[20](1,2) Judea Pearl. *Causality*. Cambridge university press, 2009.

[[21]] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI Ethics*, 1(4):529–544, 2021.

[[22]] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. Fairness in credit scoring: assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094, 2022.

[[23]]

Susan Wei and Marc Niethammer. The fairness-accuracy pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2020.

[[24]] Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: fairness versus accuracy. *arXiv preprint arXiv:2112.09975*, 2021. URL: https://arxiv.org/abs/2112.09975.

[[25]] Alice Xiang. Being'seen'vs.'mis-seen': tensions between privacy and fairness in computer vision. *Harvard Journal of Law & Technology, Forthcoming*, 2022.

[[26]] Andrew Chester, Yun Sing Koh, Jörg Wicker, Quan Sun, and Junjae Lee. Balancing utility and fairness against privacy in medical data. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1226–1233. IEEE, 2020.

[[27]] John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

[[28]] Matthias Kuppler, Christoph Kern, Ruben L Bach, and Frauke Kreuter. Distributive justice and fairness metrics in automated decision-making: how much overlap is there? *arXiv preprint arXiv:2105.01441*, 2021. URL: https://arxiv.org/abs/2105.01441.

[[29]] Faisal Kamiran, Indre Zliobaite, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Inf. Syst.*, 35(3):613–644, 2013.

This entry was written by Resmi Ramachandranpillai, Fredrik Heintz, Stefan Buijsman, Miguel Couceiro, Guilherme Alves, Karima Makhlouf, and Sami Zhioua.

[[1]] \(TP,FP,FN,\) and \(TN\) stand for: true positives, false positives, false negatives, and true negatives, respectively.

[[2]] Called explanatory features in [29].

Fair Machine Learning

In brief

Fair Machine Learning models take into account the issues of bias and fairness. Approaches can be categorized as pre-processig, which transform the input data, as in-processing, which modify the learning algorithm, and post-processing, which alter models' internals or their decisions.

More in Detail

Fairness can be promoted in three different ways in ML as surveyed in [3]. This survey provides a clear categorization of methods under pre-process, in-process and post-process approaches.

Pre-process approaches are the most flexible ones that transforms the data so that the underlying bias is removed. One advantage of using pre-process techniques is that it is the most inspectable method, as it is the earliest opportunity to mitigate biases and measure how it affects the outcome compared to the other two approaches in fair machine learning [4]. Suppression or Fairness Through Unawareness is a baseline method that accounts for removing the sensitive features and proxy sensitive features from the dataset [5]. A recent study [6] proved that removing sensitive information does not guarantee fair outcomes. Massaging the dataset (relabeling) method can act in two ways:

1. Identify unfair outcomes and correct them by changing the label to what ought to have happened.
2. Identify sensitive classes and relabel them so that the outcome is fair.

Reweighting approach has several positive aspects compared to suppression and relabeling. It works by postulating that a fair dataset would have no conditional dependence on the outcome of any of the sensitive attributes. That means it corrects the past unfair outcomes by giving more weightage to correct cases and less weightage to incorrect cases. Learning fair representations approach fairness fundamentally differently by aiming for a middle ground between-group fairness and individual fairness [7]. It turns the pre-process problem into a combined optimization problem that finds trade-offs between-group fairness, individual fairness, and accuracy.

In-process approaches modify the learning algorithms to remove biases during model training by either incorporating fairness into the optimization equation or imposing a constraint as regularization [8, 9]. The main categories of in-processing approaches are adversarial debiasing and prejudice removal. The former involves an adversary to predict the sensitive attributes from a downstream task (classification or regression), and thus the model learns a representation independent of sensitive features. Learning fair representations can be done by adding noise to the predictive power using the regulation. Adversarial reweighted learning [10] uses non-sensitive features and labels to measure unfairness and co-train the adversarial reweighting approach to improving learning. On the other hand, the prejudice remover approach has various techniques to mitigate biases during training. Some of the standard methods are:

1. Heuristic-based: Use Roony rules [11], which effectively rank problems.
2. Algorithmic Changes: These can be made in every single step of calibration, such as input, output, and model structure [4, 12, 13, 14, 15, 16, 17, 18, 19].
3. Using pre-trained models: It involves combining available pre-trained models and transferring them to reduce bias [20]
4. Counterfactual and Causal Reasoning: This considers a model to be group or individual fair if its prediction in the real world is similar to the counterfactual world, where individuals belong to a different protected group. Causal reasoning can be used to caution against those counterfactual explanations [21] [22]. A primary concern on the use and misuse of counterfactual fairness has been studied in [23].

Finally, post-process approaches are the most versatile approaches if the model is already in the production stage and it does not require retraining the model. Another advantage of using post-processing is that the fairness (individual and group) of any downstream tasks can be easily satisfied concerning the domain and application of the model [24]. Also, post-processing is agnostic to the input data, which makes it easier to implement. However, post-processing procedures may present weaker results when compare to pre-processing ones [22, 25].

Assessment tools: Tools can assist practitioners or organizations in documenting the measures, providing guidance, helping formalize processes, and empowering automated decisions. There are various types of tools to identify and mitigate the biases. Out of which, technical/quantitative tools and qualitative tools are primarily used in real-world applications by engineers and data scientists. Technical/quantitative tools focus on data or AI pipeline through technical solutions. One major drawback is that it may miss essential fairness considerations; for example, it cannot be employed to mitigate bias in the COMPAS algorithm as the nuances could not be adequately captured. It lacks methods to understand and mitigate biases but perpetuates a misleading notion that “Fair ML” is not a complex task to achieve. Some of the standard solutions in this category are:

1. IBM’s AI Fairness 360 Toolkit: It is a python toolkit through the lens of technical solutions under fairness metrics.
2. Google’s What-If Tool explores the model’s performance on a dataset through hypothetical situations. It allows users to explore different definitions of fairness constraints under various feature intersections.
3. Microsoft’s fairlean.py: It is a python package consisting of mitigation algorithms and metrics for model assessment.

On the other hand, Quantitative techniques can delve into the nuances of fairness. They can enable teams to explore the societal implications, analyses fairness harms and tradeoffs, and propose plans to find the potential sources of bias and ways to mitigate them. Two of the most prominent qualitative techniques are:

1. Co-designed AI fairness checklist (2020): This checklist is designed by a group of Microsoft researchers and academicians, 49 individuals from 12 technical organizations. It covers the items included in different stages of the AI pipeline, including envision, define, prototype, build, launch, and evolve, and is customizable according to the deployment.
2. Fairness Analytic (2019): This analytic tool is developed by Mulligan et al. to promote fairness at the earlier stages of product development. It enables teams to understand biases from a specific application perspective to analyze and document their effects.

While these tools exist to analyze the potential harms, it is the responsibility of users to understand the after-effects of which tools they are using, and which types of biases can mitigate. A detailed review of landscape and gaps in fairness tool kits is given in [1].

Bibliography

[[1]] Michelle Seng Ah Lee and Jatinder Singh. The landscape and gaps in open source fairness toolkits. In *CHI*, 699:1–699:13. ACM, 2021.

[[2]] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. Fairness in credit scoring: assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094, 2022.

[[3]] Simon Caton and Christian Haas. Fairness in machine learning: a survey. *arXiv preprint arXiv:2010.04053*, 2020. URL: https://arxiv.org/abs/2010.04053.

[4](1,2) Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. Conscientious classification: a data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134, 2017.

[[5]] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017. URL: https://arxiv.org/abs/1710.03184.

[[6]] Boris Ruf and Marcin Detyniecki. Active fairness instead of unawareness. *arXiv preprint arXiv:2009.06251*, 2020. URL: https://arxiv.org/abs/2009.06251.

[[7]] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML (3)*, volume 28 of JMLR Workshop and Conference Proceedings, 325–333. JMLR.org, 2013.

[[8]] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, 63(4/5):4:1–4:15, 2019.

[[9]] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *CoRR*, 2017.

[[10]] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*. 2020.

[[11]] Caitlin Kuhlman, MaryAnn Van Valkenburg, and Elke A. Rundensteiner. FARE: diagnostics for fair ranking using pairwise error metrics. In *WWW*, 2936–2942. ACM, 2019.

[[12]] Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. A confidence-based approach for balancing fairness and accuracy. In *SDM*, 144–152. SIAM, 2016.

[[13]] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *KDD*, 2212–2220. ACM, 2019.

[[14]] Dylan Slack, Sorelle A. Friedler, and Emile Givental. Fairness warnings and fair-MAML: learning fairly with minimal data. In *FAT**, 200–209. ACM, 2020.

[[15]] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. Putting fairness principles into practice: challenges, metrics, and improvements. In *AIES*, 453–459. ACM, 2019.

[[16]] Jialu Wang, Yang Liu, and Caleb C. Levy. Fair classification with group-dependent label noise. In *FAccT*, 526–536. ACM, 2021.

[[17]] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *FAT*, volume 81 of Proceedings of Machine Learning Research, 119–133. PMLR, 2018.

[[18]] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In *ICLR*. 2016.

[[19]] Anay Mehrotra and L. Elisa Celis. Mitigating bias in set selection with noisy protected attributes. In *FAccT*, 237–248. ACM, 2021.

[[20]] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *ICML*, volume 80 of Proceedings of Machine Learning Research, 3381–3390. PMLR, 2018.

[[21]] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018. URL: https://arxiv.org/abs/1805.05859.

[[22]] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *ICML*, volume 97 of Proceedings of Machine Learning Research, 4674–4682. PMLR, 2019.

[[23]] Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *FAccT*, 228–236. ACM, 2021.

[[24]] Pranay Kr. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *ICASSP*, 2847–2851. IEEE, 2019.

[[25]] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

This entry was written by Resmi Ramachandran Pillai, Fredrik Heintz, Miguel Couceiro, and Guilherme Alves.

Grounds of Discrimination

In brief

International and national laws prohibit **discriminating on some explicitly defined grounds**, such as race, sex, religion, etc. They can be considered in isolation, or interacting, giving rise to multiple discrimination and intersectional discrimination.

More in Detail

The Universal Declaration of Human Rights prohibit discrimination in several grounds [1]: 1) race, 2) skin colour, 3) sex, 4) language, 5) religion, 6) political or other opinion, 7) national or social origin, 8) property, or 9) birth [2], although the list is not exhaustive. By directly addressing these grounds, the Declaration highlights the problematic of considering decisions or regulations on them while leaves the door open to a more extensive view by prohibiting as well discrimination based on *other grounds*. By doing so, the Declaration implies that any difference in treatment or exercise with respect to the rights encompassed in the Declaration would have legal implications. Therefore, grounds of discrimination should not be considered a closed and fixed list but an enumeration opened to debate and reflection as the circumstances and context require. For example, the African Charter on Human and People's Rights prohibits discrimination in grounds of fortune, rather than property [3] whereas the American Convention on Human Rights includes economic status [4] and the Charter of Fundamental Rights of the European Union adds the association with a national minority [4].

To this regard, *grounds of discrimination* encompass three different motives on which decisions and policies should not be based (see also the entry Discrimination & Equity): (1) grounds innate to the individual such as race, gender, age, disability, (2) grounds intrinsic to the individual freedom and autonomy that is political belief or religion, and (3) grounds highly founded on stereotypes or stigma and which are usually irrelevant for social, economic, or politic interactions, as sexual orientation or ethnicity) [6]. The use of any of these grounds is often perceived as a lack of impartiality influenced by negative and prejudiced reasons and emotions towards certain members of the society. Prohibiting discrimination on these grounds aims to ensure that the distribution of social goods and services do not respond to subjective and irrational feelings, whether that turns out to be an advantage or a disadvantage for the individual or group concerned.

Sex refers to a person's biological status, categorized as male, female, or intersex; *gender* refers to the attitudes and behaviors that a culture associates with a person's sex, categorized as masculine, feminine and transgender (gender identity different from sex assigned at birth or non-binary); *sexual orientation* refers to the sex of those to whom one is sexually and romantically attracted, categorized as homosexual, heterosexual, and bisexual. See [7] for a psychological discussion of the differences between the terms, [8] for a discussion with reference to the United States (U.S.) anti-discrimination law, and [9] for a comparative analysis on anti-discrimination European Law. A country profile report on the legal rights of lesbian, gay, bisexual and transgender and (LGBT) people is published yearly[2] by Human Rights Watch. Human-Computer Interaction research is also addressing the extent to which AI systems "express gender and sexuality explicitly and through relation of experience and emotions, mimicking the human language on which they are trained" [10].

Race is a social construct to categorize people into groups. The term is controversial, and with little consensus on its actual meaning. [11] summarizes biological and social concepts of race, and discuss U.S. categorizations of races used for data collection, e.g., in census data. *Ethnicity* refers to self-identifying groups based on beliefs concerning shared culture, ancestry and history. The distinction between *race* and *ethnic* grounds is, nonetheless, a provocative issue primarily in Europe where after the Second War World the notion of *race* become some sort of a *taboo*. By consequence, the lacking of words, academic work, and policies addressing *race (un)justice* has also resulted on a downplay of race grounds of discrimination and the indistinct use of *ethnic origin* as *race* with mislead intentions [12].

Legislations and research studies have evolved with a different focus on vulnerable groups, sometimes restricting themselves to specific settings, including credit and insurance, sale, rental, and financing of housing, personnel selection and wages, access to public accommodation, and education. For instance, discrimination against Afro-Americans is dealt with to a large extent by studies from the U.S., whilst discrimination against Roma people has been mainly considered by E.U. studies.

Although the aforementioned grounds for discrimination are typically considered separately, the interaction of multiple forms of discrimination has been receiving increasing attention [13, 14]. An elderly disabled woman for example, could be discriminated against for being above a certain age, because she is a woman, because she is disabled, or any combination of these. *Multiple* discrimination comes into play when a person is discriminated against on the basis of different characteristics *at different times*: each type of discrimination works independently, according to distinct experiences, and multiple discrimination refers to their cumulative impact. When different grounds operate *at the same time*, then this is known as *compound* or *intersectional* discrimination. Compound discrimination (sometimes called *additive multiple discrimination*) occurs when each ground *adds* to discrimination on other grounds, for example migrant women experiencing both under-employment (such as migrants compared to local residents) and lower pay (such as female workers compared to male workers). Intersectional discrimination occurs when concurrent acts of discrimination result in a specific and distinct form of discrimination [15]. For example, [16] reports the case of Afro-American women stereotypes which when taken in isolation cover neither women nor Afro-Americans.

Grounds of discrimination are key inputs in the design of fair AI systems: fairness metrics, for instance, rely on comparing models' performances across protected and unprotected groups. We refer to the entries on and for details. Here, we concentrate on the problem of faithfully representing grounds of discrimination in data, by distinguishing the coding of human identity in raw data (*datafication*) and the representativeness of grounds of discrimination in data (*representation bias*).

For instance, if gender is coded with a binary feature (male/female), then any further discrimination analysis is limited to contrasting only such two groups, excluding non-binary people. There is then the need for a more elaborate representation of human identity in raw data, e.g, using ontologies for concept reasoning [17]. Moreover, the categories used to encode grounds of discrimination may embed forms of structural discrimination, which is hidden when features are considered in isolation, but made apparent when connected with other features in a knowledge graph [18]. The issue of *source criticism* [19], which is central historical and humanistic disciplines, is still in its infancy in the area of big data and AI. Source criticism attains at the provenance, authenticity, and completeness of data collected, especially in social media platforms. For instance, the mechanisms of social software, such as the option given to users to identify their gender as binary, result into functional biases [20] of the data collected. Beyond the complexity of datafication, the representiveness of grounds of discrimination in datasets [21] also affects discrimination and diversity analyses, and the fairness of AI models trained over those datasets (see also the entry on Bias).

Most of the grounds of discrimination fall in the category of sensitive personal data whose collection and processing is prohibited under several privacy and data protection laws, unless certain exceptions apply. For example, the grounds of race, ethnic origin, sexual orientation, political stances, religious beliefs, and trade union membership are considered special categories of personal data under the European General Data Protection Regulation [1]. Likewise, the California Privacy Rights Act (CPRA) will include as sensitive attributes, among other, consumer's racial or ethnic origin, religious or philosophical beliefs [22], while the Virginia Consumer Data Protection Act (VCDPA) will add to those attributes, the ground of mental or physical health, sexual orientation, or citizenship or immigration status [23]. From a regulatory perspective, the restriction towards the collection and processing of sensitive personal data intends to minimize the possibilities of algorithmic systems to discriminate people based on intrinsic or innate attributes of the individual. However, some criticism have arisen towards this perspective as more voices defend the need to use sensitive attributes to ensure the non-discriminatory nature of algorithmic models [24]. The European Proposal for Regulating Artificial Intelligence (Artificial Intelligence Act) seems to have reflect on this position as it will introduce a exception allowing, to the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the processing of special categories of data [25].

Discrimination grounds in datasets can be be the output of an inference. For instance, gender may be explicitly given (e.g., in a registration form) with consent to a specific usage (e.g., personalization), or it can be inferred using supervised learning [26]. A growing number of AI approaches can infer people's personality traits [27], to be used e.g., for personalization and recommendation purposes. To some extent, even in cases where the system is blinded to protected attributes, inferences can lead to discriminatory results as the system finds correlations directly related to grounds of discriminatory. Inferences can, therefore, be quite problematic because they can reinforce the historical disadvantage and inequalities suffered by certain members of the society [28]. As the current legal protections rest on the restricted access to data reveling the belonging of an individual to a protected group or prohibition of use of such data to motivate a decision, the access and use of such information indirectly creates a threat to individuals' rights [29]. For this reason, the correctness of such inferences can be crucial on attributed grounds of discrimination and, consequently, on decisions and fairness analyses. Despite inferences offer new possibilities for biased and invasive decision-making, the legal status of inferred personal, both with respect to data protection and anti-discrimination laws, is quite debated [30].

Bibliography

[[1]] European Parliament & Council. General Data Protection Regulation. 2016. L119, 4/5/2016, p. 1–88.

[[2]] UN General Assembly and others. Universal declaration of human rights. *UN General Assembly*, 302(2):14–25, 1948.

[[3]] The African Union. African Charter on Human and Peoples' Rights. 1981.

[[4]] Organization of American States. American Convention on Human Rights Pact of San Jose, Costa Rica (B-32). 1969.

[[5]] "European Parliament and the Council". Charter of fundamental rights of the european union. 2007.

[[6]] Janneke Gerards. The discrimination grounds of article 14 of the european convention on human rights. *Human Rights Law Review*, 13(1):99–124, 2013.

[[7]] American Psychological Association. Guidelines for psychological practice with lesbian, gay, and bisexual clients. 2011. URL: http://www.apa.org/pi/lgbt/resources/guidelines.aspx.

[[8]] Mary Anne C. Case. Disaggregating gender from sex and sexual orientation: The effeminate man in the law and feminist jurisprudence. *The Yale Law Journal*, 105(1):1–105, 1995.

[[9]] FRA. Protection against discrimination on grounds of sexual orientation, gender identity and sex characteristics in the EU-comparative legal analysis. 2015.

[[10]] Justin Edwards, Leigh Clark, and Allison Perrone. Lgbtq-ai? exploring expressions of gender and sexual orientation in chatbots. In *CUI*, 2:1–2:4. ACM, 2021.

[[11]] Rebecca M. Blank, Marilyn Dabady, and Constance F. Citro, editors. *Measuring Racial Discrimination - Panel on Methods for Assessing Discrimination*. National Academies Press, 2004.

[[12]] Nicolas Kayser-Bril. Europeans can't talk about racist ai systems. they lack the words. 2021. URL: https://algorithmwatch.org/en/europeans-cant-talk-about-racist-ai-systems-they-lack-the-words/.

[[13]] European Commission. Tackling multiple discrimination: Practices, policies and laws. 2007. Directorate General for Employment, Social Affairs and Equal Opportunities, Unit G.4. URL: http://ec.europa.eu/social/main.jsp?catId=738\&pubId=51.

[[14]] ENAR. European network against racism, Fact sheet 44: the legal implications of multiple discrimination. 2011. URL: https://www.enar-eu.org/wp-content/uploads/fs44_-_the_legal_implications_of_multiple_discrimination_final_en.pdf.

[[15]] European Commission, Directorate-General for Justice and Consumers and Sandra Fredman. *Intersectional discrimination in EU gender equality and non-discrimination law*. Publications Office, 2016.

[[16]] T. Makkonen. Compound and intersectional discrimination: bringing the experiences of the most marginalized to the fore. 2002. Unpublished manuscript, Institute for Human Rights, Abo Alademi University.

[[17]] Clair A. Kronk and Judith W. Dexheimer. Development of the gender, sex, and sexual orientation ontology: evaluation and workflow. *J. Am. Medical Informatics Assoc.*, 27(7):1110–1115, 2020.

[[18]] Christopher L. Dancy and P. Khalil Saucier. AI and blackness: towards moving beyond bias and representation. *CoRR*, 2021.

[[19]] Gertraud Koch and Katharina Kinder-Kurlanda. Source criticism of data platform logics on the internet. *Historical Social Research*, 45(3):270–287, 2020.

[[20]] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: biases, methodological pitfalls, and ethical boundaries. *Frontiers Big Data*, 2:13, 2019.

[[21]] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. A survey on techniques for identifying and resolving representation bias in data. *CoRR*, 2022.

[[22]] California Statu Legislature and the Council. California consumer privacy act of 2018 [1798.100 - 1798.199.100]. 2018.

[[23]] Virginia Senate. Virginia consumer data protection act. Effective January 1, 2023.

[[24]] Indrė Žliobaitė and Bart Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2):183–201, 2016.

[[25]] European Parliament and the Council. Regulation of the european parliament and of the council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. 2021.

[[26]] Lucía Santamaría and Helena Mihaljevic. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput. Sci.*, 4:e156, 2018.

[[27]] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Trans. Affect. Comput.*, 5(3):273–291, 2014.

[[28]]

Raphaele Xenidis. Tuning eu equality law to algorithmic discrimination: three pathways to resilience. *Maastricht Journal of European and Comparative Law*, 27(6):736–758, 2020.

[[29]] Solon Barocas. Data mining and the discourse on discrimination. In *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining*, 1–4. 2014.

[[30]] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Columbia Business Law Review*, 2019(2):494–620, 2019.

This entry was written by Alejandra Bringas Colmenarejo and Salvatore Ruggieri.

[[1]] Protected group, protected grounds and prohibited grounds are also used as synonymous of grounds of discrimination.

[[2]] For 2021, see the Human Right Watch World Report.

Justice

In brief

Justice encompasses three different perspectives: (1) *fairness* understood as the fair treatment of people, (2) *rightness* as the quality of being fair or reasonable, and (3) a legal system, the scheme or system of law. Justice can be distinguished between *substantive* and *procedural*.

More in Detail

It is commonly accepted that, *justice* entails “the proper administration of the law; the fair and equitable treatment of all individuals under the law” [1]. Therefore, *justice* encompasses three different perspectives, (1) *fairness* understood as the fair treatment of people, (2) *rightness* as the quality of being fair or reasonable, and (3) a legal system, the scheme or system of law, in which every person receives his/her/its due from the system, including all rights, both natural and legal [5]. Artificial Intelligence (AI) can be a tool for administering justice in the legal system, or it can itself be subject to the requirements of fairness and rightness when used for automated decision making (ADM). In the former case, AI can be adopted at several levels of autonomy [6], e.g., from no automation to superhuman autonomous AI for legal reasoning. For a state of the art of the use of Machine Learning (ML) in the criminal justice system (mainly in the United States), see [7]. Several books and newspapers commentaries warn about the risks of using AI for justice administration [8]. In the latter case, the design of AI-based systems can benefit from discussion and theories of justice in the legal and ethical disciplines. However, the above conceptualization of *justice* has given rise to an endless and ongoing debate regarding whether justice is an inherent component of the law, not separate or distinct from it, or is simply a moral judgment about law [9]. In essence, the debate considers whether justice understood as fairness and rightness is independent from the law, or to what extent the *law* includes considerations of justice and the legal system simply applies justice to human conflicts. In conclusion, the concept of *justice* is as central to legal theory as it is difficult to define.

Nevertheless, a range of different components or categories of justice have been defined, both in the philosophical literature [10] and in the legal literature [11]. These can be understood along several distinctions, starting with one between *substantive* and *procedural* justice. This is the difference between considering justice in terms of the outcomes which have to meet certain standard in order to be just [12], versus considering justice in terms of a procedure which meets certain standards (and possibly considering the outcomes of any such procedure as being just regardless of the resulting distributions [13]). It is, however, common to consider procedural justice partly in terms of the results (e.g., a trial procedure is just if it – at least – mostly acquits the innocent and punishes the guilty). As such, substantive justice is the main notion to discuss here, although the use of AI in procedures also affects questions of procedural justice.

Substantive justice in turn can be viewed in different ways. First, there is a question of whether one focuses on *distributive* justice or on *corrective* justice. Distributive justice deals with the distribution of the benefits and burdens of social cooperation [12]. These can be *comparative*, such as the theory of (strict) egalitarianism which requires that resources are distributed to minimize overall inequality [14] and other versions of egalitarianism. For example, on Luck egalitarianism (and, closely related, Equality of Opportunity) inequalities in the final distribution may be allowed only in so far as they are not the result of luck or a difference of opportunity [15, 16]. By far the most influential, however, has been Rawls' view of Justice as Fairness, which combines a requirement of equality of opportunity with the difference principle: unequal distributions have to satisfy a min-max condition where "they are to be to the greatest benefit of the least advantaged members of society." [17] Alternatively, distributive notions of justice can be *non-comparative*. Sufficiency principles [18], requiring that everyone receives a minimally sufficient amount of resources are a clear example of a distributive justice notion that doesn't involve comparisons between individuals. Desert-based principles [19], which hold that resources should be allocated based on what individuals deserve can also be non-comparative (if they specify absolute amounts based on what one deserves, as opposed to a share of the total). Views thus differ on what the right principles are for distributive justice. Furthermore, it is interesting that most of these principles have only a limited overlap to *fairness* in ADM [3] (cf. the entry on Fairness notions and metrics).

Where distributive justice focuses on the just distribution of goods, corrective justice concerns the rectification of wrongs or the undoing of transactions which can be either voluntary (contract) or involuntary (when defrauded or a victim of misrepresentation) [20, 21]. This differs from distributive justice, as corrective justice first requires a wrong that needs to be corrected, and the correction might violate the ideal distribution of goods according to distributive justice principles. As such, disagreements exist over the priority to be placed between these two principles: is corrective justice merely a way to achieve distributive justice or is corrective justice normatively prior? [22] However this issue is settled, it is a matter of fact that corrective justice is an important part of current legal systems. Similarly, *retributive* justice [23], which focuses on the compensation of the victim of criminal behaviour and the punishment of the lawbreaker, is crucial to our current systems.

In all these cases procedures are followed to make decisions on the distribution of resources, the appropriate corrections and potential punishments. *Procedural* justice relates to the normative conditions that these procedures have to meet. As such, it encompasses principles of legality, proportionality, effective remedy, fair trial, presumption of innocence and right of defence [4, 24, 25, 26]. Procedural justice is also affected by the use of AI, as this changes procedures and so new standards have to be found for when a procedure including AI is just. Such standards are also needed for transparency, and the notion of procedural justice has been used to propose such a standard by [27], who argue that what matters to determine the justness of an algorithm is the goals of the algorithm as well as how effectively they are met. That is intended to allow an evaluation of the justness of the procedure, and thus of the use of the algorithm. For instance, [28] conducted experiments involving laypeople, that showed a "fairness gap" between human judges and AI robot judges. Such a gap is reduced by enhancing the interpretability of AI decisions.

Part of the question of what procedures are just is that of which political procedures should be decided on. *Political* justice addresses the foundational issues of political rights and responsibilities embedded in constitutional theory and how individuals shall share the control over the shape of the constitution[1] [29]. *Social* justice, on the other hand, addresses how members should compare under the basic structure of the society [29], and, its "primary task is not so much to save the computational infrastructure AI and ICTs rely on but rather to defend society" [30]. This concern decisions about the broad shape of society and thus cannot readily be solved with fair ML tools. Yet algorithms can help in the implementation of these decisions. As such, *justice* can be seen to differ from *fairness*: its scope is often broader, and it is not restricted to questions of equality between different groups to which an algorithm is applied. See e.g., [31] for a discussion of *data justice*, pertaining to "the way people are made visible, represented and treated as a result of their production of digital data", and the literature on organizational justice theory [32, 33] for the notion of *interactional justice* pertaining to how workers are treated with respect and dignity (*interpersonal justice*) and how they are provided with explanations of business process and outcomes (*informational justice*). Still, the design of AI systems intersects with all of the notions of justice discussed here.

The increased use of AI and ADM reflects a tendency to *solutionism* [34], where technical solutions are offered to solve all social and economic problems [35]. However, unfair, discriminatory, and unjustified decisions affecting different aspects of individuals' economy and private life has encouraged a critical reflection on questions regarding "what, then, do we talk about when we talk about governing algorithms?" [36]. Likewise, the proposed *algorithmic justice* [37] strives to address these unintentional effects provoked by the use of ADM for the allocation of welfare services. This novel conception of justice, founded on Nancy Fraser's *abnormal justice theory*, defends "the need to expand our collective understanding of justice, beyond issues of equal access to, and equal distribution of justice” [38, 39] as referred in [37] in order to recognise, debate, diagnose and address harmful effects of ADM in the allocation of transformative services [37]. That being the case, (un)just ADM are more and more scrutinised under the lens of *procedural justice* as several studies [40, 41, 42] “suggest that people do not only care about whether the outcome of a decision benefits them, but also whether it meets standards of justice” [2]. To this regard, AI’s *explainability and transparency* would be crucial to justify and explain the algorithmic decisions and decision-making process and therefore ensure the accountability intelligibility of the decisions, and, by extension, of the process (a principle known as *open justice* when referring to the judicial system [43]).

Bibliography

[[1]] Jeffrey Lehman, Shirelle Phelps, and others. *West's encyclopedia of American law*. Thomson/Gale, 2004.

[[2]] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In *CHI*, 377. ACM, 2018.

[[3]] Matthias Kuppler, Christoph Kern, Ruben L Bach, and Frauke Kreuter. Distributive justice and fairness metrics in automated decision-making: how much overlap is there? *arXiv preprint arXiv:2105.01441*, 2021. URL: https://arxiv.org/abs/2105.01441.

[[4]] "European Parliament and the Council". Charter of fundamental rights of the european union. 2007.

[[5]] Gerald N Hill and Kathleen Hill. *The people's law dictionary: Taking the mystery out of legal language*. MJF Books, 2002.

[[6]] Lance Eliot. Identifying a set of autonomous levels for AI-based computational legal reasoning. *MIT Computational Law Report*, 2021.

[[7]] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[[8]] Katherine B. Forrest. *When Machines Can Be Judge, Jury, and Executioner: Justice the the Age of Artificial Intelligence*. World Scientific, 2021.

[[9]] Anthony D'Amato. On the connection between law and justice. *UC Davis L. Rev.*, 26:527, 1992.

[[10]] David Miller. Justice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[[11]] Richard Susskind. *Online Courts and the Future of Justice*. Oxford University Press, 2019.

[12](1,2) John Rawls. *A theory of justice*. Harvard university press, 2020.

[[13]] Robert Nozick. *Anarchy, state, and utopia*. Volume 5038. new york: Basic Books, 1974.

[[14]] Iwao Hirose. *Egalitarianism*. Routledge, 2014.

[[15]] Elizabeth S Anderson. What is the point of equality? *Ethics*, 109(2):287–337, 1999.

[[16]] Jonathan Wolff. Fairness, respect and the egalitarian ethos revisited. *The Journal of Ethics*, 14(3):335–350, 2010.

[[17]] John Rawls. Political liberalism. *The John Dewey essays in philosophy*, 1993.

[[18]] Gillian Brock. Sufficiency and needs-based approaches. *The Oxford Handbook of Distributive Justice*, pages 86–108, 2018.

[[19]] Jeffrey Moriarty. Desert-based justice. In *The Oxford handbook of distributive justice*, pages 152–175. Oxford University Press New York, 2018.

[[20]] Arthur Ripstein. The division of responsibility and the law of tort. *Fordham L. Rev.*, 72:1811, 2003.

[[21]] Ernest J Weinrib. *The idea of private law*. Oxford University Press, 2012.

[[22]] Steven Walt. Eliminating corrective justice. *Va. L. Rev.*, 92:1311, 2006.

[[23]] Alec Walen. Retributive Justice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

[[24]] Steven L Blader and Tom R Tyler. A four-component model of procedural justice: defining the meaning of a "fair" process. *Personality and social psychology bulletin*, 29(6):747–758, 2003.

[[25]] Denise Meyerson and Catriona Mackenzie. Procedural justice and the law. *Philosophy Compass*, 13(12):e12548, 2018.

[[26]] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural justice in algorithmic fairness: leveraging transparency and outcome control for fair algorithmic mediation. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):182:1–182:26, 2019.

[[27]] Michele Loi, Andrea Ferrario, and Eleonora Viganò. Transparency as design publicity: explaining and justifying inscrutable algorithms. *Ethics and Information Technology*, 23(3):253–263, 2021.

[[28]] Benjamin Minhao Chen, Alexander Stremitzer, and Kevin Tobia. Having your day in robot court. *Harvard Journal of Law & Technology*, 2022. Forthcoming. URL: https://ssrn.com/abstract=3841534.

[29](1,2) David Sobel, Peter Vallentyne, and Steven Wall. *Oxford Studies in Political Philosophy, Volume 1*. Oxford University Press, 2015.

[[30]] Jasmina Tacheva, Sepideh Namvarrad, and Najla Almissalati. A higher purpose: towards a social justice informatics research framework. In *iConference (1)*, volume 13192 of Lecture Notes in Computer Science, 265–271. Springer, 2022.

[[31]] Linnet Taylor. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, 2017.

[[32]] Sarah Bankins, Paul Formosa, Yannick Griep, and Deborah Richards. Decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context. *Information Systems Frontiers*, 2022.

[[33]] Lionel P. Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Hum. Comput. Interact.*, 35(5-6):545–575, 2020.

[[34]] Evgeny Morozov. *To save everything, click here : technology, solutionism and the urge to fix problems that don't exist*. London : Allen Lane, 2013.

[[35]] Aleš Završnik. Algorithmic justice: algorithms and big data in criminal justice settings. *European Journal of criminology*, 18(5):623–642, 2021.

[[36]] Solon Barocas, Sophie Hood, and Malte Ziewitz. Governing algorithms: A provocation piece. 2013. URL: https://ssrn.com/abstract=2245322.

[37](1,2,3) Olivera Marjanovic, Dubravka Cecez-Kecmanovic, and Richard Vidgen. Theorising algorithmic justice. *European Journal of Information Systems*, pages 1–19, 2021.

[[38]] Nancy Fraser. Abnormal justice. *Critical inquiry*, 34(3):393–422, 2008.

[[39]] Nancy Fraser. Injustice at intersecting scales: on 'social exclusion'and the 'global poor'. *European journal of social theory*, 13(3):363–371, 2010.

[[40]] E Allan Lind and Tom R Tyler. *The social psychology of procedural justice*. Springer Science & Business Media, 1988.

[[41]] Jason A Colquitt and Jessica B Rodell. Measuring justice and fairness. In *The Oxford handbook of justice in the workplace*, pages 187–202. Oxford University Press, 2015.

[[42]] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *ICHI*, 160–169. IEEE Computer Society, 2015.

[[43]] Adam Pah, David Schwartz, Sarath Sanga, Charlotte Alexander, Kristian Hammond, Luis Amaral, and SCALES OKN Consortium. The promise of AI in an open justice system. *AI Magazine*, 43(1):69–74, 2022.

This entry was written by Alejandra Bringas Colmenarejo, Stefan Buijsman, and Salvatore Ruggieri.

[[1]] See the European Social Survey page for a survey on perceptions of political justice in Europe.

Segregation

In brief

Social segregation refers to the separation of groups on the grounds of personal or cultural traits. Separation can be physical (e.g., in schools or neighborhoods) or virtual (e.g., in social networks).

More in Detail

Social segregation refers to the “*separation of socially defined groups*” [1]. People are partitioned into two or more groups on the grounds of personal or cultural traits that can foster discrimination, such as gender, age, ethnicity, income, skin color, language, religion, political opinion, membership of a national minority, etc. (see entry on Grounds of Discrimination). Contact, communication, or interaction among groups are limited by their physical, working or socio-economic distance. Such a separation is observed when dissecting the society into organizational units (neighborhoods, schools, job types).

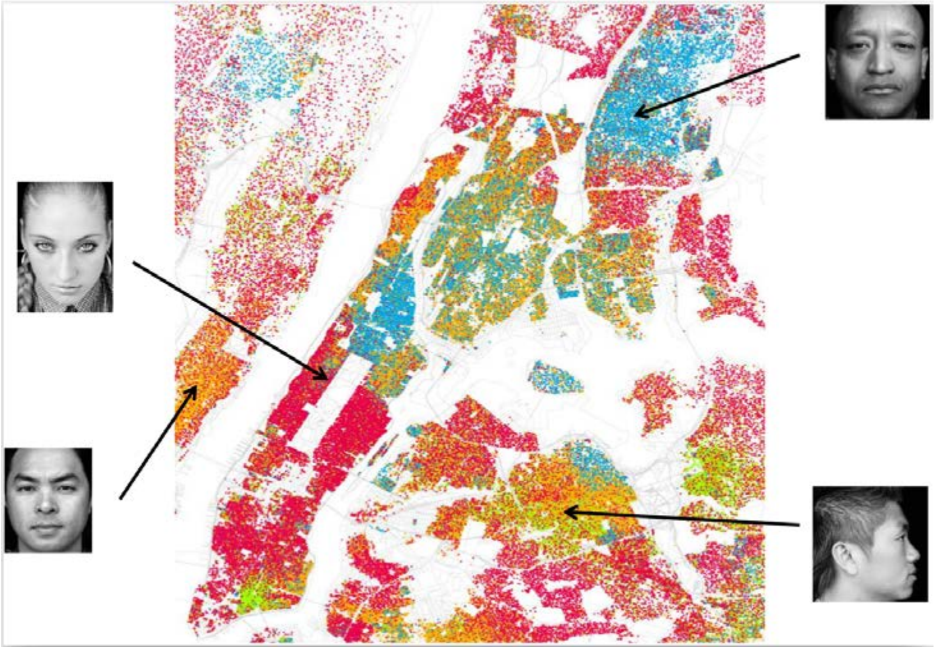


Fig. 17 Racial spatial segregation in New York City, based on Census 2000 data [2]. One dot for each 500 residents. Red dots are Whites, blue dots are Blacks, green dots are Asian, orange dots are Hispanic, and yellow dots are other races.

Early studies on residential segregation trace back to 1930's [3]. In this context, social groups are set apart in neighborhoods where they live in, in schools they attend to, or in companies they work at. As sharply pointed out in Fig. 17, racial segregation (a.k.a. residential segregation on the grounds of race) very often emerges in cities characterized by ethnic diversity. Schelling's segregation model [4, 5] shows that there is a natural tendency to spatial segregation, as a collective phenomenon, even if each individual is relatively tolerant – in his famous abstract simulation model, Nobel laureate Schelling assumed that a person changes residence only if less than 30% of the neighbors are of his/her own race.

[6] argued that segregation is shifting from ancient forms on the grounds of racial, ethnic and gender traits to modern socio-economic and cultural segregation on the basis of income, job position, and political-religious opinions. An earlier comparison of ideological segregation of the American electorate online and offline is offered in [7]. The paper found that segregation in news consumption is higher online than offline, but significantly lower than the segregation of face-to-face interactions with neighbors, co-workers, or family members. More recently, it has been warned that the filter bubble generated by personalization of online social networks may foster segregation [8], opinion polarization [9], and lack of consensus between different social groups. Segregation in social network has been investigated in [10], with experiments on segregation on the grounds of sex and age for directors in the boards of the companies. Other works have focused on religious social networks [11],

A segregation index provides a quantitative measure of the degree of segregation of social groups (e.g., Blacks, Whites, Hispanics, etc.) distributed among units of social organization (e.g., schools, neighborhoods, jobs, etc.). Several indexes have been proposed in the literature. The surveys [12, 13] represent the earliest attempts to categorize them. Afterward, [14] provided a shared classification with reference to five key dimensions: evenness, exposure, concentration, centralization, and clustering. Finally, [15] adapts segregation measure to graphs representing social networks. In this entry, we will consider basic evenness and exposure indexes. Other three classes of indexes are specifically concerned with spatial notions of segregation. Concentration indexes measure the relative amount of physical space occupied by social groups in an urban area. Centralization indexes measure the degree to which a group is spatially located near the center of an urban area. Clustering indexes measure the degree to which group members live disproportionately in contiguous areas.

We restrict here to consider binary indexes, which assume a partitioning of the population into two groups, say majority and minority (but could be men/women, native/immigrant, White/NonWhite, etc.). Let T be size of the total population, $0 < M < T$ be the size of the minority group, and $P = M/T$ be the overall fraction of the minority group. Assume that there are n organizational units (or simply, units), and that for $i \in [1, n]$, t_i is the size of the population in unit i, m_i is the size of the minority group in unit i, and $p_i = m_i/t_i$ is the fraction of the minority population in unit i.

Evenness indexes. Evenness indexes measure the difference in the distributions of social groups among organizational units. The *dissimilarity index* D is the weighted mean absolute deviation of every unit's minority proportion from the global minority proportion:

$$D = \frac{1}{2 \cdot P \cdot (1-P)} \sum_{i=1}^n \frac{t_i}{T} \cdot | p_i - P | \label{equ:dissimilarity}$$

The normalization factor $2 \cdot P \cdot (1-P)$ is to obtain an index in the range $[0, 1]$. Since D measures dispersion of minorities over the units, higher values of the index mean higher segregation. Dissimilarity is minimum when for all $i \in [1, n]$, $p_i = P$, namely the distribution of the minority group is uniform over units. It is maximum when for all $i \in [1, n]$, either $p_i = 1$ or $p_i = 0$, namely every unit includes members of only one group (complete segregation).

The second widely adopted index is the *information index*, also known as the *Theil index* in social sciences [16] and normalized mutual information in machine learning [17]. Let the population entropy be $E = - P \cdot \log{P}-(1-P) \cdot \log{(1-P)}$, and the entropy of unit i be $E_i = - p_i \cdot \log{p_i}-(1-p_i) \cdot \log{(1-p_i)}$. The information index is the weighted mean fractional deviation of every unit's entropy from the population entropy:

$$H = \sum_{i=1}^n \frac{t_i}{T} \cdot \frac{(E-E_i)}{E}$$

Information index ranges in $[0, 1]$. Since it denotes a relative reduction in uncertainty in the distribution of groups after considering units, higher values mean higher segregation of groups over the units. Information index reaches the minimum when all the units respect the global entropy (full integration), and the maximum when every unit contains only one group (complete segregation).

The third evenness measure is the *Gini index*, defined as the mean absolute difference between minority proportions weighted across all pairs of units, and normalized to the maximum weighted mean difference. In formula:

$\label{eq:Gini} G = \frac{1}{2 \cdot T^2 \cdot P \cdot (1-P)} \cdot \sum_{i=1}^n \sum_{j=1}^n t_i \cdot t_j \cdot |p_i - p_j|$
Here $\sum_{i=1}^n \sum_{j=1}^n t_i \cdot t_j \cdot |p_i - p_j|$ is the weighted mean absolute difference. The normalization factor is obtained by maximizing such a value. The definition of the Gini index stems from econometrics, where it is used as a measure of the inequality of income distribution [18]. The Gini index ranges in $[0, 1]$ with higher values denoting higher segregation. The maximum and minimum values are reached in the same cases of the dissimilarity index.

Exposure indexes. Exposure indexes measure the degree of potential contact, or possibility of interaction, between members of social groups. The most used measure of exposure is the *isolation index* [19], defined as the likelihood that a member of the minority group is exposed to another member of the same group in a unit. For a unit i, this can be estimated as the product of the likelihood that a member of the minority group is in the unit (m_i/M) by the likelihood that she is exposed to another minority member in the unit (m_i/t_i, or p_i) – assuming that the two events are independent. In formula:
$I = \frac{1}{M} \cdot \sum_{i=1}^n m_i \cdot p_i$

The right hand-side formula can be read as the minority-weighted average of minority proportions in units. The isolation index ranges over $[P, 1]$, with higher values denoting higher segregation. The minimum value is reached when for $i \in [1, n]$, $p_i = P$, namely the distribution of the minority group is uniform over the units. The maximum value is reached when there is only one $k \in [1, n]$ such that $m_k = t_k = M$, namely there is a unit containing all minority members and no majority member.

A dual measure is the *interaction index*, which is the likelihood that a member of the minority group is exposed to a member of the majority group in a unit. By reasoning as above, this leads to the formula:
$\mathit{Int} = \frac{1}{M} \cdot \sum_{i=1}^n m_i \cdot (1-p_i)$
It clearly holds that $I + \mathit{Int} = 1$. Hence, lower values denote higher segregation. A more general definition of interaction index occurs when more than two groups are considered in the analysis, so that the exposure of the minority group to one of the other groups is worth to be considered [14].

The key problem of assessing social segregation has been investigated by hypothesis testing, i.e., by formulating one or more possible contexts of segregation against a certain social group, and then in empirically testing such hypotheses. Such an approach is currently supported by statistical tools, such as the R packages *OasisR*[1] and *seg*[2] [20], or by GIS tools such as the *Geo-Segregation Analyzer*[3] [21]. A tool for multidimensional exploration of segregation index has been proposed[4] in [22].

Bibliography

[[1]] Douglas S Massey. Segregation and the perpetuation of disadvantage. In *The Oxford Handbook of the Social Science of Poverty*, pages 369–393. Oxford University Press, 2016.

[[2]] Eric Fischer. Distribution of race and ethnicity in US major cities. 2011. under Creative Commons licence, CC BY-SA 2.0. URL: http://www.flickr.com/photos/walkingsf/sets/72157624812674967/detail/.

[[3]] Nancy A Denton and Douglas S Massey. Residential segregation of Blacks, Hispanics, and Asians by socioeconomic status and generation. *Social Science Quarterly*, 69(4):797–817, 1988.

[[4]] Thomas C Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2):143–186, 1971.

[[5]] W. A. V. Clark. Residential preferences and neighborhood racial segregation: a test of the Schelling segregation model. *Demography*, 28(1):1–19, 1991.

[[6]] Douglas S. Massey, Jonathan Rothwell, and Thurston Domina. The changing bases of segregation in the United States. *Annals of the American Academy of Political and Social Science*, 626:74–90, 2009.

[[7]] Matthew Gentzkow and Jesse M Shapiro. Ideological segregation online and offline. *Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

[[8]]

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80:298–320, 2016. URL: http://ssrn.com/abstract=2363701.

[[9]] Michael Maes and Lukas Bischofberger. Will the personalization of online social networks foster opinion polarization? 2015. URL: http://ssrn.com/abstract=2553436.

[[10]] Alessandro Baroni and Salvatore Ruggieri. Segregation discovery in a social network of companies. *J. Intell. Inf. Syst.*, 51(1):71–96, 2018.

[[11]] Jiantao Hu, Qian-Ming Zhang, and Tao Zhou. Segregation in religion networks. *EPJ Data Sci.*, 8(1):6:1–6:11, 2019.

[[12]] Otis Dudley Duncan and Beverly Duncan. A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210–217, 1955.

[[13]] D. R. James and K. E. Tauber. Measures of segregation. *Sociological Methodology*, 13:1–32, 1985.

[14](1,2) Douglas S Massey and Nancy A Denton. The dimensions of residential segregation. *Social Forces*, 67(2):281–315, 1988.

[[15]] Michal Bojanowski and Rense Corten. Measuring segregation in social networks. *Soc. Networks*, 39:14–32, 2014.

[[16]] R. Mora and J. Ruiz-Castillo. Entropy-based segregation indices. *Sociological Methodology*, 41:159–194, 2011.

[[17]] T. Mitchell. *Machine Learning*. The Mc-Graw-Hill Companies, Inc., 1997.

[[18]] Joseph L Gastwirth. A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society*, 39(6):1037–1039, 1971.

[[19]] Wendell Bell. A probability model for the measurement of ecological segregation. *Social Forces*, 32(4):357–364, 1954.

[[20]] Seong-Yun Hong, David O'Sullivan, and Yukio Sadahiro. Implementing spatial segregation measures in R. *PLoS ONE*, 9(11):e113767, 2014.

[[21]] Philippe Apparicio, Joan Carles Martori, Amber L. Pearson, Éric Fournier, and Denis Apparicio. An open-source software for calculating indices of urban residential segregation. *Social Science Computer Review*, 32(1):117–128, 2014.

[[22]] Alessandro Baroni and Salvatore Ruggieri. Scube: A tool for segregation discovery. In *EDBT*, 542–545. OpenProceedings.org, 2019.

This entry was readapted from *Alessandro Baroni and Salvatore Ruggieri. Segregation discovery in a social network of companies. J. Intell. Inf. Syst., 51(1):71–96, 2018* by Salvatore Ruggieri

[[1]] cran.r-project.org/package=OasisR

[[2]] cran.r-project.org/package=seg

[[3]] geoseganalyzer.ucs.inrs.ca

[[4]] github.com/ruggieris/SCube

Accountability and Reproducibility

In brief

Accountability and Reproducibility are two interrelated concepts, cornerstones of Trustworthy AI. Accoutable AI systems can contribute to reproducibility, and Reproducible AI systems can contribute to accountability.

More in detail

Accountability and Reproducibility are two cornerstones of Trustworthy AI [1]. Accountability requires mechanisms be put in place to ensure that AI systems and their outcomes, both before and after their development, deployment and use, can be observed and analyzed. This ability to review AI systems involve technical and organisational logging processes [3] to enable investigators to draw the same conclusions from an experiment by following provided guidelines.

In this context, Accountability and Reproducibility are interrelated concepts. Developing reprodubicle AI systems can enable accountability over AI systems. On the other hand, the process of record-tracking and logging for accountability can support an increasing level of reproducibility.

A third dimension strictly correlated with Accountability and Reproducibility is Traceability. We suggest to navigate in the appropriate section of this book for more detailed information about these three dimensions.

Main Keywords

- Accountability: **Accountability** is an ethical aspect studied in the TAILOR project to ensure that a given actor or actors can render an account of the actions of an AI system. The accountability concept is strictly related to the concept of responsibility.
- Wicked problems: A class of problems for which science provides insufficient or inappropriate resolution.
- Meaningful human control: **Meaningful human control** is the notion that aims to generalize the traditional concept of operational control over technological artifacts to artificial intelligent systems. It implies that artificial systems should not make morally consequential decisions on their own, without appropriate control from responsible humans.
- The Frame Problem: The **frame problem** is the challenge of knowing and modeling the relevant features and context of situations, and getting an agent to act on those without consideration all the irrelevant facts as well.
- Reproducibility: **Reproducibility** is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.
- Traceability: **Traceability** can be defined as the need to maintain a complete and clear documentation of the data, processes, artefacts and actors involved in the entire lifecycle of an AI model, starting from its design and ending with its production serving.
- Provenance Tracking: **Provenance tracking** represents the tracking of “information that describes the production process of an end product, which can be anything from a piece of data to a physical object. […] Essentially, provenance can be seen as meta-data that, instead of describing data, describes a production process.”
- Continuous Performance Monitoring: **Continuous performance monitoring** is the activity to track, log and monitor over time the behaviour and the performance of Artificial Intelligence and Machine Learning models. This activity is particularly relevant after in-production deployment in order to detect any performance drifts and outages of the model.

Bibliography

[[1]] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Reviewable automated decision-making: a framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. 2021.

[[2]] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

This entry was written by Luciano C Siebert.

Accountability

In Brief

Accountability is an ethical aspect studied in the TAILOR project to ensure that a given actor or actors can render an account of the actions of an AI system. The accountability concept is strictly related to the concept of responsibility.

Abstract

According to [1], the requirement of accountability complements the other ethical dimensions, and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.

Motivation and Background

Whenever something goes wrong, there is often a call to define who is responsible for this wrongdoing. Responsibility is a broader topic that might have different conceptualizations. However, in this sense, it usually means one's obligation to render an account of your actions and the consequences of these, i.e. accountability. Accountability can be defined as a form of *"passive responsibility"* (or backward looking responsibility) in the sense of being held to account for or justify towards others a given action or consequence that happened in the past [4]. Although accountability implies having to account for one's actions, if the account given is considered insufficient, then one might still be considered blameworthy and thus deserving of censure or blame [5].

AI systems bring particular concerns with respect to accountability, as understanding how the systems work can be challenging, and commercial considerations can conceal broader organisations processes [3]. Although one might "understand" the inner workings of the algorithms used, the outcomes might still not be predictable, complicating accountability even further [6].

Two often discussed examples to explain the accountability setting are related to autonomous driving cars and medical decisions.

Indeed, imagine a self-driving car that hits a pedestrian. Who should account for or justify the system's actions? The person within the car that might not have been able or willing to supervise the system? The manufacturer, that designed the systems and thus should be the only responsible of the behaviour of its products? The programmer that did not correctly implement all the necessary checks? The manufacture of the sensor that did not detect the pedestrian? The person who conducted the test that did not foresee that particular circumstance?

Or, again, in the case of a wrong diagnosis following an MRI. Do we expect an account from the doctor that did not see the error or the AI system (at all the same possible levels we saw in the previous example)?

Given the difficulties, a lot of effort put in the definition of accountability regards the **Auditability** principle: [1]

> Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited.

Other aspects took into consideration in the High Level Expert Group report [1] are:

- **Minimisation and reporting of negative impacts**, i.e., assessing, documenting and minimising the potential negative impacts of AI systems, even thanks to the use of impact assessments both prior to and during the development, deployment and use of AI systems.
- **Trade-offs** to tackle tensions that may arise between requirements. If conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights. Any decision about which trade-off to make should be reasoned and properly documented. Whether

no ethically acceptable trade-offs can be identified, [1] clearly states that the development, deployment and use of the AI system should not proceed in that form.

- **Redress** must be ensured when things go wrong, with particular attention to vulnerable persons or groups. The importance of the redress is advocated also by the European Union Agency for Fundamental Rights [7], where particular emphasis is posed to collective redress (i.e., collective redress, a way in which victims can join forces to overcome obstacles), and by the Council of Europe [8], where is specified that a citizen should not necessarily have to pursue legal action straight away and seeking remedies should be available, known, accessible, affordable and capable of providing appropriate redress.

Guidelines

Several guidelines and checklists have been proposed to increase accountability over the actions of AI systems, both from EU authorities and industries, such as:

- **The Assessment List for Trustworthy Artificial Intelligence (ALTAI)** [9]
  This Assessment List (ALTAI) is firmly grounded in the protection of people's fundamental rights exposed by the High Level Expert Group report [1]. It is probably the most complete one so far and the reference point for all other checklists.
  The ALTAI checklist helps organisations understand what Trustworthy AI is, in particular what risks an AI system might generate, and how to minimize those risks while maximising the benefit of AI. It is intended to help organisations identify how proposed AI systems might generate risks, and to identify whether and what kind of active measures may need to be taken to avoid and minimise those risks. It aims at raising awareness of the potential impact of AI on society, the environment, consumers, workers and citizens (in particular children and people belonging to marginalised groups) and at encouraging the multidisciplinarity and the involvement of all relevant stakeholders. It helps to gain insight on whether meaningful and appropriate solutions or processes to accomplish adherence to the seven requirements (as outlined above) are already in place or need to be put in place. This could be achieved through internal guidelines, governance processes, etc.
  For each requirement, this Assessment List for Trustworthy AI (ALTAI) provides introductory guidance and relevant definitions in the Glossary. The online version of this assessment list contains additional explanatory notes for many of the questions.
- **Getting the Future Right** [10]
  The European Union Agency for Fundamental Rights published a report where a fundamental rights-based analysis of concrete 'use cases' is provided. The report illustrates some of the ways that companies and the public sector in the EU are looking to use AI to support their work, and whether – and how – they are taking fundamental rights considerations into account. In this way, it contributes empirical evidence, analysed from a fundamental rights perspective, that can inform EU and national policymaking efforts to regulate the use of AI tools.
- **ICO's guidance on the use of artificial intelligence** [2, 11]
  The UK data protection authority has been very active on all the topics related to accountability, publishing guidance that is constantly updated. In [2], the focus of accountability is on being compliant with data protection law and being capable of minimise risks. It explores some important aspects such as Leadership and oversight (e.g., the structure of the analyzed organization), Transparency, Privacy, and Security; a whole section is dedicated to the Data Protection Impact Assessment.
  In [11], a short checklist is presented, even if in the document itself is highlighted that "Accountability is not a box-ticking exercise" but rather taking responsibility for what you are doing with personal data, considering this as an opportunity to develop and sustain people's trust.
- **IBM's FactSheets** [12]
  This document starts from Supplier's Declarations of Conformity (SDoCs), which are documents largely used by many industries even if they are usually not legally required documents, to describe the lineage of a product along with the safety and performance testing it has undergone. SDoCs aims at capturing and quantifying various aspects of the product and its development to make it worthy of consumers' trust. The chechlist proposed in [12] should help increasing trust in AI services. We envision such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers.
  A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance (including appropriate accuracy or risk measures along with timing information), safety,

explainability, algorithmic fairness, and security and robustness. Moreover, the FactSheet should help in listing how the service was created, trained, and deployed along with what scenarios it was tested on, how it may respond to untested scenarios, guidelines that specify what tasks it should and should not be used for, and any ethical concerns of its use. Hence, FactSheets help prevent overgeneralization and unintended use of AI services by solidly grounding them with metrics and usage scenarios. FactSheet is a quite interesting example because in this case a private company highlights the need of ethical procedures, standards, and certifications.

- **Microsoft's guideline for human-AI interaction** [13]
  In this paper, authors identified 18 question, related to different phases of the use of an AI system, and 10 different kinds of application, ranging from e-commerce recommender systems to route planning systems, from automatic photo organizers to social network feed filtering systems. Then, authors empirically evaluated both the clarity and the relevance of various questions in the various domains, highlighting potential criticality (e.g., reporting a violation to the question ``Make clear why the system did what it did'' if a recommender system did not non give any explanations of the reason why a certain product was suggested).

Possible Taxonomy of terms

Boven [14] defines accountability as

> a **relationship** between an **actor** and a **forum**, in which the actor has an obligation to **explain and to justify** his or her conduct, the forum can pose questions and pass judgement, and the actor may face **consequences**.

Wiering [15] presented a thorough systematic literature review on algorithmic accountability structured on the five points identified by Boven in his definition, which we briefly summarize below:

- **Arguments on the actor**: Involves a broader discussion on who is responsible for the harm that the system may inflict when it is working correctly, and who is responsible when it is working incorrectly. It involves different levels of actors (e.g. individuals, teams, department, organizations) with different roles and possibly also third-parties. Due to these multiple levels and actors, situations known as ``the problem of many hands'' might occur, in which the collective can reasonably be held responsible for an outcome, while none of the individuals can be reasonably held responsible for that outcome [4]. In these situations, to be "in the loop" is not enough, calling for a more meaningful ability to control the design and operation process, in other words calling for L3.meaningful_human_control.
- **The forum**: To whom a given account is directed. The forum might take different shapes such as political, legal, administrative, professional, and towards the civil society. Examples of forum include General Data Protection Regulation (GDPR), the proposed EU AI Act, or even the guidelines for trustworthy AI proposed by the European Commission [1], as discussed in the previous section.
- **The relationship between the actor and the forum**: This relationship comes in different forms and shapes, according to all other four points. Nevertheless, they are usually mapped in three phases: the information phase, the deliberation and discussion phase, and the final phase, where consequences can be imposed on the actor by the forum.
- **The content and criteria of the account**: Although ex ante analysis, such as impact assessment and simulations, can be helpful, they are limited as they cannot foresee all possible behavior and consequences. The importance of an accountability relationship should depend not only on such ex ante factors, but also on the extent to which a given system impacts society and individuals.
- **The consequences which may result from the account**: In situations where there is a more "vertical" accountability relationship between actor and forum (e.g., accountability through legal standers), consequences are usually made more tangible. In more "horizontal" settings (e.g., self-regulation of organizations), consequences are defined based more on a moral imperative.

Accountability gaps and their implications

As AI systems, especially systems with learning abilities, are deployed "in the wild", human control and prediction over their behaviour are very difficult if not impossible, leading to so-called *"accountability gaps"*. Santoni de Sio & Mecacci [16] divided such gaps in public accountability gap and moral accountability gaps. *Public* *accountability gaps* relate to citizens not being able to get an explanation for decisions taken by public agencies, while *moral accountability gap* refers to the reduction of human agents' capacity to make sense of – and explain to each other the behaviour of AI systems.

Accountability gaps point to the fact that accounting requires knowledge and some ability to control [16]. Designers and developers of AI systems can only tackle this challenge by acknowledging that this is not a matter of fortuitous allocation of praise or blame, and that systems should be developed in a manner that allows for stakeholders to be held accountable. Among other things, this relates to the social context where these systems are deployed (and whether a given solution or formulation can be argued), how the questions and criteria for accountability are framed, and the level of understanding and control that a given actor might have. In this encyclopedia, we discuss these three interrelated concepts, respectively, in the entries Wicked problems, The Frame Problem, and Meaningful human control.

Bibliography

[1](1,2,3,4,5,6) High Level Expert Group on AI. Ethics Guidelines for Trustworthy AI. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on 2022-05-10).

[2](1,2) Information Commissioner's Office (ICO). Accountability framework. URL: https://ico.org.uk/for-organisations/accountability-framework/ (visited on 2022-05-25).

[[3]] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. Reviewable automated decision-making: a framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. 2021.

[4](1,2) Ibo R Van de Poel and Lambèr MM Royakkers. *Ethics, technology, and engineering: An introduction*. Wiley-Blackwell, 2011.

[[5]] Ibo van de Poel. The relation between forward-looking and backward-looking responsibility. In *Moral responsibility*, pages 37–52. Springer, 2011.

[[6]] Marijn Janssen and George Kuk. The challenges and limits of big data algorithms in technocratic governance. 2016.

[[7]] European Union Agency for Fundamental Rights. Improving access to remedy in the area of business and human rights at the EU leve. 2016. URL: https://fra.europa.eu/en/opinion/2017/business-human-rights (visited on 2022-05-10).

[[8]] Council of Europe. Guide to human rights for internet users: effective remedies and redress. 2014. URL: http://www.coe.int/en/web/internet-users-rights/guide (visited on 2022-05-10).

[[9]] High Level Expert Group on AI. The Assessment List for Trustworthy Artificial Intelligence (ALTAI). 2020. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (visited on 2022-05-10).

[[10]] European Union Agency for Fundamental Rights. Getting the future right - Artificial Intelligence and Fundamental Rights. 2020. URL: https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights (visited on 2022-05-10).

[11](1,2) Information Commissioner's Office (ICO). Accountability and governance. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/ (visited on 2022-05-25).

[12](1,2) Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. 2019. arXiv:1808.07261v2.

[[13]]

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *ACM International Conference of Human-Computer Interaction (CHI)*. 2019.

[[14]] Mark Bovens. Analysing and assessing accountability: a conceptual framework 1. *European law journal*, 13(4):447–468, 2007.

[[15]] Maranke Wieringa. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 1–18. 2020.

[16](1,2) Filippo Santoni de Sio and Giulio Mecacci. Four responsibility gaps with artificial intelligence: why they matter and how to address them. *Philosophy & Technology*, pages 1–28, 2021.

This entry was written by Luciano C Siebert and Francesca Pratesi.

Wicked problems

In brief

A class of problems for which science provides insufficient or inappropriate resolution [1].

More in Detail

Wicked problems are not objectively given but their formulation already depends on the viewpoint of those presenting them [2].

In spatial planning literature there is a difference between tame problems and wicked problems. The former is a problem with a set of well-defined rules and clear goal, e.g. problems like solving sudoku’s. There are however another set of problems that do not do well when we think of them in terms of search spaces, constraints, rules, and goal settings.

This class of problems, named Wicked problems [1] is largely determined by the professional skill of framing and addressing the problem in a particular way. These problems are political like policy around poverty. The setting and solutions are contingent, depending on political view, available information, and dependent on formulation. There are ten different markers that show wickedness[1].

1. There is no definite formulation of a wicked problem.
2. Wicked problems have no stopping rule.
3. Solutions to wicked problems are not true-or-false, but good-or-bad.
4. There is no immediate test of a solution to a wicked problem.
5. Every solution to a wicked problem is a ’one-shot operation’; because there is no opportunity to learn by trial-and-error, every attempt counts significantly.
6. Wicked problems do not have an enumerable set of potential solutions, nor is there a well-described set of permissible operations that may be incorporated into the plan.
7. Every wicked problem is essentially unique.
8. Every wicked problem can be considered to be a symptom of another problem.
9. The existence of a discrepancy representing a wicked problem can be explained in numerous ways. The choice of explanation determines the nature of the problem’s resolution.
10. The planner has no right to be wrong

Spatial planning may be far removed at first sight from engineering, but with the embedding of technology in society, we do move towards a society were tame problems with well-defined rules fall short with regard to the potential impact implementations may have in society. In fact, many algorithms can already be regarded as policy in one way other another, as a particular implementation may provide benefits to a certain train of thought or a particular set of actions. A policy that gives financial benefits those in dire straits (e.g. those below the poverty line) can have a similar effect as an algorithm that allocates resources to those in need. Point being, we should not underestimate the similarity between administration and algorithm.

If we take the ten different markers for algorithmic implementation into account, then we can draw a number of inferences. First, it shows us the immediate impact of implementation, not only is implementation a one-shot operation because it may skew public perception[3], it also may influence other potential solutions. Second, these solutions are political and not to be framed in terms of optimization (true/false and good/bad). Third, there are a variety of equally effective solutions that would be permissible, meaning that the choice for this particular one carries a certain political or personal weight.

The first point latches onto something else in planning theory, that of path dependency. When an implementation becomes embedded in society it is hard to remove[4]. We also see this in philosophy technology through the Collingride dilemma[5]. In planning theory it shows that implementation may effect future implementations as it opens certain doors and closes others. Consider for example, our use of the QWERTY keyboard, this is partly due to its widespread use, rather than efficiency (DVORAK is more effective) [6]. The claim of path dependency is that these implementations (of which many others would be equivalent) can cause a path that is hard to step away from.

This last inference is one that engineers should take into account when designing algorithms for a societal context. It means that they themselves become a political player in the scheme of things rather than the executioner of the wishes of certain stake-holders. In essence, the role of the engineers and that of a policy designer are interlinked by their societal impact. Wicked problems are a way of showing the impact they have when dealing with bureaucracy, algorithmic or otherwise. Accountability comes into play when we consider that the engineer is not a neutral player within this game, they carry some of the blame of the outcome as it was their framing of the problem that led to a particular solution. Of course there are many ways to alleviate some of these problems

Bibliography

[1](1,2,3) Horst WJ Rittel and Melvin M Webber. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169, 1973.

[[2]] Richard Coyne. Wicked problems revisited. *Design studies*, 26(1):5–17, 2005.

[[3]] Bas Verplanken. Beliefs, attitudes, and intentions toward nuclear energy before and after chernobyl in a longitudinal within-subjects design. *Environment and Behavior*, 21(4):371–392, 1989.

[[4]] Richard J Lazarus. Super wicked problems and climate change: restraining the present to liberate the future. *Cornell L. Rev.*, 94:1153, 2008.

[[5]] David Collingridge. *The social control of technology*. St. Martin's Press, New York, 1982.

[[6]] Paul A David. Clio and the economics of qwerty. *The American economic review*, 75(2):332–337, 1985.

This entry was written by Sietze Kuilman.

Meaningful human control

In brief

Meaningful human control is the notion that aims to generalize the traditional concept of operational control over technological artifacts to artificial intelligent systems. It implies that artificial systems should not make morally consequential decisions on their own, without appropriate control from responsible humans.

More in Detail

The notion of meaningful human control has its origins in the discussions on lethal autonomous weapon systems (LAWS), specifically in regards to life-or-death decisions that such systems could in principle make. Avoiding ethical issues related to autonomous decision making by artificial agents requires that humans, and only humans, have control of and are accountable for the use of lethal force [1]. The concrete implications of this requirement are still debated, with proposals range from calls for a full ban of LAWS [2] to suggestions on governance, implementation, and use of such systems that can contribute to meaningful human control (e.g. [3], [4]). While ethical issues associated with the lack of human control are perhaps most apparent for autonomous weapon systems, they extend far beyond the military domain, to a wider class of human-AI systems that make decisions with moral implications. At the time of writing, researchers have approached meaningful human control in the contexts of automated driving systems [5] [6], medical decision support systems [7], unmanned aerial vehicles [8], among other domains. Many of these domain-specific operationalizations rely on a philosophical account of meaningful human control proposed by [9]. This account builds on the concept of “guidance control” [10] and provides two necessary conditions for meaningful human control. The tracking condition requires that the decision-making system tracks and responds to all human reasons (i.e., values, norms, intentions) relevant in given circumstances. The tracing condition requires that any action/decision of the human-AI system should be traceable to at least one human within the system who has proper moral understanding of the situation and the effects of the system in that situation.

Tracking and tracing, as well as several alternative domain-specific accounts, provide conceptual frameworks for meaningful human control. Making these concepts less vague and more relatable to design and engineering practice is however very challenging [11]. In [12] an attempt to close this gap between theory and practice is made by proposing four actionable properties that can be addressed throughout the system’s lifecycle:

- Property 1: A system in which humans and AI algorithms interact should have an explicitly defined domain of morally loaded situations within which the system ought to operate.
- Property 2: Humans and AI agents within the system should have appropriate and mutually compatible representations.
- Property 3: Responsibility attributed to a human should be commensurate with that human’s ability and authority to control the system.
- Property 4: There should be explicit links between the actions of the AI agents and actions of humans who are aware of their moral responsibility.

Bibliography

[[1]] missing institution in article36_2014

[[2]] missing institution in article36_2015

[[3]] Heather M Roff and Richard Moyes. Meaningful human control, artificial intelligence and autonomous weapons. In *Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Au-Tonomous Weapons Systems, UN Convention on Certain Conventional Weapons*. 2016.

[[4]] missing institution in horowitz2015

[[5]] Daniël D Heikoop, Marjan Hagenzieker, Giulio Mecacci, Simeon Calvert, Filippo Santoni De Sio, and Bart van Arem. Human behaviour with automated driving systems: a quantitative framework for meaningful human control. *Theoretical issues in ergonomics science*, 20(6):711–730, 2019.

[[6]] Simeon C Calvert, Bart van Arem, Daniël D Heikoop, Marjan Hagenzieker, Giulio Mecacci, and Filippo Santoni de Sio. Gaps in the control of automated vehicles on roads. *IEEE intelligent transportation systems magazine*, 13(4):146–153, 2020.

[[7]] Matthias Braun, Patrik Hummel, Susanne Beck, and Peter Dabrock. Primer on an ethics of ai-based decision support systems in the clinic. *Journal of medical ethics*, 47(12):e3–e3, 2021.

[[8]] Marc Steen, Jurriaan van Diggelen, Tjerk Timan, and Nanda van der Stap. Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives. *AI and Ethics*, pages 1–13, 2022.

[[9]]

Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: a philosophical account. *Frontiers in Robotics and AI*, pages 15, 2018.

[[10]] John Martin Fischer and Mark Ravizza. *Responsibility and control: A theory of moral responsibility*. Cambridge university press, 1998.

[[11]] Rebecca Crootof. A meaningful floor for meaningful human control. *Temp. Int'l & Comp. LJ*, 30:53, 2016.

[[12]] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M Jonker, and others. Meaningful human control: actionable properties for ai system development. *AI and Ethics*, pages 1–15, 2022.

This entry was written by Arkady Zgonnikov and Luciano C Siebert.

The Frame Problem

In brief

The **frame problem** is the challenge of knowing and modeling the relevant features and context of situations, and getting an agent to act on those without consideration all the irrelevant facts as well.

More in Detail

The frame problem originated with McCarthy and Hayes[1] back in the sixties, but it has been appropriated many times over. The frame problem, as described by McCarthy and Hayes, was mostly about representationalism. They wondered how we could describe an update function such that it does not require a multitude of unnecessary and unaffected statements. It was and is a poignant question. We could have an agent that is able of acting on certain parts of the world, say paint pieces of paper [2], but when we give it other actions, these actions may interact. How does the machine know that the moving said paper won't also change its colour? In representationalism, it seemed to mean that we had to add a variety of statements that only worked in serious edge-cases.

The frame problem as philosophers appropriated it was about generalized action. How does an agent keep a faithful representation of the world, after it has acted [3]? Such an agent would need to have a kind of update function that does not require going over all the superfluous statements [4]. Fodor [5] posited it as Hamlet's problem: How does an agent know when to stop thinking?

The frame problem these days is sometimes regarded as the general relevance problem, not being limited to representationalism but going into connectionism as well [6]. In this case, the inheritance of the frame problem for connectionism entails the follow: how does an agent know which data is considered to be relevant to the situation? The problem for connectionists approaches is not that input needs to exert influence on the system, but rather that it produces the correct influence to preserve relevance.

This is where one can see problems for accountability on the horizon. In situations where agents have to act, we require that they indeed make the correct inferences given a certain context, have the correct update function, and produce the correct influence. All of these require that the agent understands the context at hand and is able to make the correct inferences such that the relevant action is achieved.

However, that is easier said than done. The task of the agent's designer is finding a way that all these relevant inferences can be incorporated into the agent. For if they don't then we introduce a gap. The model of the agent, their capacity for action, and the world will result in an agent that acts while missing (relevant) inferences. This can thus end up harming people or misaligning with human intention. The obvious examples are those of harmful classifications - Google for example had a problem with the classification of Gorilla's[7].

The question of the frame problem for accountability is not how to solve the frame problem, as that seems to require solving a variety of tractability question and understanding the relation between agent and the world such that relevancy can be aptly captured. Rather, the designer should be aware of its own limitation and the limitations of the model that is housed within the agent. Meaning that accountability has a social aspect of explaining the necessary limits of the system or boxing possible actions of the agents such that these limits are acceptable in their scope.

Bibliography

[[1]] John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.

[[2]] missing journal in shanahan2004frame

[[3]] Daniel C Dennett. Cognitive wheels: the frame problem of ai. *Minds, machines and evolution*, pages 129–151, 1984.

[[4]] Drew McDermott. Ai, logic, and the frame problem. In *The frame problem in artificial intelligence*, pages 105–118. Elsevier, 1987.

[[5]] missing journal in fodor1987modules

[[6]] Richard Samuels. Classical computationalism and the many problems of cognitive relevance. *Studies in History and Philosophy of Science Part A*, 41(3):280–293, 2010.

[[7]] Ludovic Righetti, Raj Madhavan, and Raja Chatila. Unintended consequences of biased robotic and artificial intelligence systems [ethical, legal, and societal issues]. *IEEE Robotics & Automation Magazine*, 26(3):11–13, 2019.

This entry was written by Sietze Kuilman and Luciano C Siebert.

Reproducibility

Synonyms: Replicability, Repeatability.

In brief

Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators [1].

Abstract

This entry firstly introduces the motivations behind reproducibility in the scientific process and, then, in artificial intelligence and machine learning. Due to the rather wide range of different meanings of reproducibility in the literature and the ambiguity of the terms, a brief review of the most important definitions is provided and discussed. In this context, we promote the most stable formulation of the definition. Practical guidelines to various standards for documenting code, technical experiment setup, and data are also discussed.

Motivation and Background

Reproducibility in science means that one can repeat or replicate the same (or sufficiently similar) experiment and obtain the same (or sufficiently similar) research results as the original scientists on the basis of their publications and descriptions. To this aim and to ease the replication, the discovered claims, methods and analyses should be described in a sufficiently detailed and transparent way. Diverse reproducibility settings have been identified in the literature, see e.g. [1] [2], but from a more general standpoint, reproducibility entails that studies are reproduced by independent researchers.

Reproducibility is an essential ingredient of the scientific method, meant to verify the published results and claims and to enable a continuous self-correcting process in scientific discoveries. Unfortunately, the rising of a so-called research replication crisis has been lately pointed out ([3]). According to several surveys, a relatively too large amount of published research results, in such disciplines as chemistry, biology, medicine and pharmacy, earth and environmental sciences, cannot be repeated. This may suggest issues with these results or at least with their good descriptions. Reproducibility in artificial intelligence (AI) and, in particular machine learning (ML), are specifically challenging. The continuously increasing complexity of new methods (often having many hyper-parameters that need specialized optimization strategies), the size of studied datasets and the use of advanced computational resources pose many difficulties for communicating the necessary results as compared to the older works. The paper [4] presents the view of some researchers (such as J. Pineau citing her interview) claiming that ML was previously more theoretically based, while it has become a more experimental science in the past decade, and many proposals of new models, in particular deep networks, come from running many experiments with the intensive use of available data. In this context, the authors ([5]) indicate growing difficulties in reproducing the work of others. Other reasons of difficulty in reproducibility include: lack of access to the same training data or differences in data distribution; mis-specification or under-specification of the model or training procedure; lack of availability of the code necessary to run the experiments, or errors in the code; under-specification of the metrics used to report results; selective reporting of results and ignorance of the danger of adaptive overfitting as well as the use of adaptation strategies embedded in the development libraries.

Nevertheless, software solutions and systems based on AI and ML are gaining momentum. Many of them are being used in high-stake applications where their decisions can have an impact on people and society, and their improper operation may cause harm. In this frame, the quest for reproducibility of such methods is even more urgent and reproducibility becomes one of the key postulates within Responsible AI or Trustworthy AI. [6] also claims that reproducibility of AI is very important for other reasons. Researchers, students and R&D engineers need to have a good understanding of new and, often quite complex, methods, reproduce them (sometimes by their own re-implementations), carefully check their correctness, examine their working conditions and limitations, as well as to verify the presented results, especially if they need to further use them in their systems often applied to complex tasks. Moreover much of AI new projects receive either public or business funds, so it should be subject to accountability and it is necessary to convince others that these projects can produce reliable results.

Terminology

In this handbook, we follow the concept of reproducibility introduced by [7]. According to this concept, which is also adopted in a number of more recent papers (e.g.,[1]; [8]; [5]), *reproducibility* refers to the ability of an independent researcher to reproduce the same, or reasonably similar results using the data and the experimental setup provided by the original authors.

Reproducibility should not be confused with other terms describing the ability to replicate the results in science, such as replicability and repeatability ([2]). *Replicability* defined in a way consistent with our understanding of reproducibility is the ability of an independent researcher to produce results that are consistent with the conclusions of the original work, using new data or different the experimental setup. The term *repeatability* appears in some references, e.g. [9] that uses a notion of reproducibility inconsistent with our definition, but should be considered to describe an ability of a researcher to repeat his/her own experimental procedures using same experimental setup and data, while achieving reasonably repeatable results that support the same conclusions.

In order to compare these reproducibility-related terms, the main conceptual dimensions need to be identified. Based on the analysis of the literature, the following dimensions can be distinguished: (i) availability of the components originally deployed in experimental workflows (i.e., data, code and analysis as considered by [5]; [10]; [11]); (ii) teams involved in the experimentation (i.e., whether or not the experiments was conducted by the same group who is running the reproducibility validation); (iii) reasons because the experiment or part of it is re-conducted (i.e., validating the repeatability of the experiment or as suggested by [1] corroborating the scientific hypothesis and theory the experiment aims to support. With respect to these conceptual dimensions, the reproducibility-related terms used in the literature can be clustered in the following way:

- Most of the literature (including [5]; [1]; [12]) refers to reproducibility as the attempt to replicate experiment as much as possible as the original one, that is by using original data, code and analysis when available. Computational reproducibility, method reproducibility, direct replication and recomputation are used in lieu of

reproducibility respectively by [10], [13], [14], [15] and [16]. [1] distinguishes the notion of reproducibility from corroborating the scientific hypotheses or theory to ground which the experiment is designed for.
- The term replicability is highlighted by [7], [8], [12], [5], where an independent team can obtain the same result using the data, which could be slightly different, and methods which they develop completely independently or change slightly. Furthermore [8]; [5] use another name – robust – for carrying out the experiments with the same data and some changes in an analysis or code implementations.
- Some works such as [9] [8]; [17]; [1] uses repeatability to indicate a weaker level of reproducibility where the replication of the experiment is achieved by the same team that provided the original experiments.

In the context of the above literature review, it is also worth clarifying the discussion of what is reproduced as a result of the above activities and how to understand the term result. In the case of AI works, [1] distinguishes between different possible results to reproduce:

- Outcome – the result of applying the model implementation for selected data (e.g., predictions - labels for test examples)
- Analysis – calculated measures or other indicators (e.g. prediction accuracy values)
- Interpretation – more general conclusions from the experiments. According to Gunderesn the last point is the most important in reproducibility, because in the scientific method certain hypotheses are tested or certain beliefs are confirmed.

Similar importance of refining the levels of reproducibility has the division proposed in [13]:

- Reproducibility of methods: the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results
- Reproducibility of results: the production of corroborating results in a new study, having used the same experimental methods
- Reproducibility of inference: the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study

The general definitions should be however made more specific whenever we apply it to contemporary artificial intelligence research, and to the sub-field of machine learning in particular. The reasons are grounded in the high complication of the modern software processing pipelines, that often depend on third-party software (frameworks, libraries), use an extended set of metaparameters that are crucial to arrive at the correct results, and require modern hardware (e.g. recent GPU cards) with it’s specific architecture and drivers. These features of AI research and applications make this field different from the general science, where reproducibility refers primarily to the careful documentation of the experimental procedure.

In AI systems, the main components of the experimental setup are software and data. The software plays the role of our experimental setup. Although depending on the specific context, hardware components may be included as well (e.g. in computer vision, robotics), most of the AI-related research is conducted on pre-recorded datasets, so we can limit our scope to the software. The other dimension is data. Together, software and data define the conceptual dimensions of the space on which the defined terms are spanned in AI. However, as we noticed earlier, AI is a very broad field, with a number of distinctive sub-fields that have specific requirements when it comes to defining the exact elements of software, and sometimes have specific requirements as to the data, such as elimination of biases or privacy issues. This motivates the introduction of guidelines or “best practices” for reproducibility, that often also include terms that define the degree to which the postulate of full reproducibility is met, usually in relation to the amount of code, technical details and data that the author shares with readers.

Guidelines

Definitions of the different reproducibility-related terms are often accompanied by badges and guidelines helping people in making the definitions operational.

- Some definitions differentiate the notion of reproducibility according to the kind of resource shared. For example [10] focus on *computational reproducibility* with **bronze, silver, gold** standards. [11] and [1] propose different increasing levels *R1, R2, R3, R4* depending on whether experiment descriptions, codes, data and experiment are stored. @ACMv1.1 recommends that three separate badges related to artefact review be associated with research articles in ACM publications: Artifacts Evaluated, Artifacts Available and Results Validated.

- Guidelines ease the description of experiments. For example, [5] provides a special Machine learning reproducibility checklist; datasheets [18], model cards [19] and factsheets [20] provides templates for describing datasets and the AI models deployed increasing the transparency and accountability of experimentations and operational intelligent systems.

Below a few of the above guidelines are precised. Following [10]'s proposal, the three degrees of the reproducibility standards for ML are based on availability of data, model, and code, as well as other analyses or programming dependencies. For instance, in the bronze standards (the minimal requirements for reproducibility) the authors should make the data, model and its source code publicly available for downloading. The silver standard extends it by additionally providing: dependencies of the analysis (in a form to be installed in a single command), recording key details of the analysis and used software requirements. Furthermore, all elements in the analysis should be documented to be set deterministic. Within the gold standard the authors should also prepare this analysis reproducible with a single command - which is the most demanding with respect to full automatization of the reproducibility process.

[5] specify the necessary elements to be documented and made public with respect to the following categories: model and algorithm, theoretical claims, datasets used in experiments, shared code including dependencies specifications, all reported experimental results (with all details for the experimental setup, hyper-parameters, training details, definitions of evaluation measures, and description of the computing infrastructure used). [8] provide similar recommendations for ML in robotics, by focusing on the reproducibility of computation experiments on real robots. They stress the role of managing properly the software dependencies, distinguishing between experimental code and library code, and documenting the measurement metrics, which is essential for reinforcement learning.

Datasheets by [18] specify how to document the motivation, composition, collection process, recommended uses for data deployed in the systems and experiments; model cards by [19] ease the description of model's intended use cases limiting their usage in contexts for which they are not well suited; factsheets [20] provide a template for describing the purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers.

Software frameworks supporting reproducibility

Lately, a paradigm based on tailoring the DevOps approach to AI and ML is emerging as a practical tool for ensuring reproducibility. This paradigm makes use of frameworks for Machine Learning Model Operationalization Management (MLOps), which streamline the whole development lifecycle of AI and ML models. MLOps enables developers and auditors to keep track of and inspect the various choices done and the artefacts produced in the different phases of AI and ML design and development (i.e., data gathering, data analysis, data transformation/preparation, model training and development, model validation, and model serving). [21] analyze some of the available open tools for MLOps. This allows for maintaining a comprehensive documentation that is at the basis of model reproducibility.

Bibliography

[1](1,2,3,4,5,6,7,8,9) Odd Erik Gundersen. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2197):20200210, 2021. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0210, arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2020.0210, doi:10.1098/rsta.2020.0210.

[2](1,2) Hans E. Plesser. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, 11:76, January 2018. URL: http://journal.frontiersin.org/article/10.3389/fninf.2017.00076/full (visited on 2021-11-18), doi:10.3389/fninf.2017.00076.

[[3]] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.

[[4]] Will Douglas Heaven. Ai is wrestling with a replication crisis. *MIT Technology Review*, 2020.

[5](1,2,3,4,5,6,7,8)

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Hugo Larochelle. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *arXiv:2003.12206 [cs, stat]*, December 2020. arXiv: 2003.12206. URL: http://arxiv.org/abs/2003.12206 (visited on 2021-09-26).

[[6]] Edward Raff. A step toward quantifying independently reproducible machine learning research. 2019. arXiv:1909.06674.

[7](1,2) Jon F. Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, 601–604. 2005.

[8](1,2,3,4,5) Nicolai A. Lynnerup, Laura Nolling, Rasmus Hasle, and John Hallam. A survey on reproducibility by evaluating deep reinforcement learning algorithms on real-world robots. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of Proceedings of Machine Learning Research, 466–489. PMLR, 30 Oct–01 Nov 2020. URL: https://proceedings.mlr.press/v100/lynnerup20a.html.

[9](1,2) ACM Artifact Review and Badging - Version 1.1. August 2020. URL: https://www.acm.org/publications/policies/artifact-review-and-badging-current (visited on 2022-01-20).

[10](1,2,3,4) Benjamin J. Heil, Michael M. Hoffman, Florian Markowetz, Su-In Lee, Casey S. Greene, and Stephanie C. Hicks. Reproducibility standards for machine learning in the life sciences. *Nature Methods*, October 2021. URL: https://www.nature.com/articles/s41592-021-01256-7 (visited on 2021-11-18), doi:10.1038/s41592-021-01256-7.

[11](1,2) Odd Erik Gundersen and Sigbjørn Kjensmo. State of the Art: Reproducibility in Artificial Intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence,*. 2018.

[12](1,2) National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC, 2019. ISBN 978-0-309-48616-3. URL: https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science, doi:10.17226/25303.

[13](1,2) Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, 2016. URL: https://www.science.org/doi/abs/10.1126/scitranslmed.aaf5027, arXiv:https://www.science.org/doi/pdf/10.1126/scitranslmed.aaf5027, doi:10.1126/scitranslmed.aaf5027.

[[14]] Stephan Guttinger. The limits of replicability. *European Journal for Philosophy of Science*, 10(2):1–17, 2020.

[[15]] Ian P Gent and Lars Kotthoff. Recomputation. org: experiences of its first year and lessons learned. In *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, 968–973. IEEE, 2014.

[[16]] Victoria C Stodden. Trust your science? open your data and code. 2011. URL: https://magazine.amstat.org/blog/2011/07/01/trust-your-science/.

[[17]] Joint Committee for Guides in Metrology. The international vocabulary of metrology – basic and general concepts and associated terms - 3rd edition with minor corrections. *JcGM*, 2012. URL: https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf.

[18](1,2) Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, March 2020. arXiv: 1803.09010. URL: http://arxiv.org/abs/1803.09010 (visited on 2021-09-14).

[19](1,2) Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. Atlanta GA USA, January 2019. ACM. URL: https://dl.acm.org/doi/10.1145/3287560.3287596 (visited on 2021-09-14), doi:10.1145/3287560.3287596.

[20](1,2) Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. *arXiv:1808.07261 [cs]*, February 2019. URL: http://arxiv.org/abs/1808.07261 (visited on 2021-09-14).

[[21]] Philipp Ruf, Manav Madan, Christoph Reich, and Djaffar Ould-Abdeslam. Demystifying mlops and presenting a recipe for the selection of open-source tools. *Applied Sciences*, 2021. URL: https://www.mdpi.com/2076-3417/11/19/8861, doi:10.3390/app11198861.

This entry was written by Riccardo Albertoni, Sara Colantonio, Piotr Skrzypczyński, and Jerzy Stefanowski.

Traceability

In Brief

Traceability can be defined as the need to maintain a complete and clear documentation of the data, processes, artefacts and actors involved in the entire lifecycle of an AI model, starting from its design and ending with its production serving [2].

Abstract

This entry introduces the motivations behind traceability and illustrates its core requirements, which encompass documenting the entire development cycle of an AI model and tracking its live functioning after the deployment in production.

Motivation and Background

Developing an Artificial Intelligence (AI) model or an AI-powered system entails a considerable number of choices along the entire development process, which may result in diverse behaviours and functioning of the same model or system. This phenomenon is particularly relevant when learning-based approaches comes into play, due to the dependency of Machine Learning (ML) models on the data used for their training as well as the complexity and variety of the ML methods that might be used, especially when based on Deep Learning (DL). Furthermore, the development of such models relies often on large trial-&-error experimental processes, which are not commonly well documented (see the reproducibility entry).

This condition makes it evident the need for a comprehensive and clear documentation of the actions taken as well as the various processing steps performed when developing an AI or ML model, as, without this documentation, it might be difficult to reconstruct the reasons behind the outcomes and the functioning of an AI model. In consideration of this, the High-Level Expert Group on AI (AI HLEG) has included the traceability of an AI model as one of the main mean to enable the transparency principle for Trustworthy AI [1]. Overall, traceability aims to ensure the avoidance of any “grey” area about the AI model or system, thus guaranteeing the transparency of and the trust in the development, production functioning and usage of an AI system. The record-keeping activity entailed by traceability should regard the data used, the data pre-processing steps as well as the development settings, the development workflows and the actors involved [3]. This encompasses the detailed provision of information about the provenance and the usage of any data and artefacts involved in the development of the AI model or system. In this view, traceability incorporates the measures to ensure reproducibility and it can be understood as the technological mean for guaranteeing the auditability and accountability of AI models and systems [4].

The two souls of traceability

AI models based on learning are data-inductive and dynamic systems, whose development relies on an initial set of data. This set, although large, might not necessarily span the whole variability range of real-world cases or conditions. This implies that, when used in practice, the AI model or system can encounter slightly different or novel data that differ from those it has been exposed to during training. This phenomenon calls for the monitoring of AI models after their deployment in production, in order to log their usage as well as to track over time their performance, vitality and conduct. Such an AI maintenance system is an important part of traceability, which can be then seen as one principle, two souls:

- provenance tracking of data, processes and artefacts involved in the development of the AI model,
- continuous performance monitoring of the AI model after deployment in production.

These two aspects will be further explained in the following entries.

Bibliography

[[1]] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[[2]] European Commission, Content Directorate-General for Communications Networks, and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office, 2020. URL: https://doi.org/10.2759/002360, doi:10.2759/002360.

[[3]] Michael Bücker, Gero Szepannek, Alicja Gosiewska, and Przemyslaw Biecek. Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1):70–90, 2022. doi:10.1080/01605682.2021.1922098.

[[4]] Karim Lekadir, Richard Osuala, Catherine Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Aussó, Leonor Cerdá Alberich, Kostas Marias, Manolis Tsiknakis, Sara Colantonio, Nickolas Papanikolaou, Zohaib Salahuddin, Henry C Woodruff, Philippe Lambin, and Luis Martí-Bonmatí. Future-ai: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. 2021. URL: https://arxiv.org/abs/2109.09658, doi:10.48550/ARXIV.2109.09658.

This entry was written by Sara Colantonio.

Provenance Tracking

In brief

Provenance Tracking represents the tracking of “information that describes the production process of an end product, which can be anything from a piece of data to a physical object. […] Essentially, provenance can be seen as meta-data that, instead of describing data, describes a production process.” [1]

More in Detail

“Traceability in AI shares part of its scope with general-purpose recommendations for provenance … ”[2]. In fact, provenance is “any information that describes the production process of an end product, which can be anything from a piece of data to a physical object” [1].

Keeping track of provenance is vital in various settings. It contributes to increasing the trust on produced systems. According to Gil et al. [3], the provenance of scientific results, i.e., “how results were derived, what parameters influenced the derivation, what datasets were used as input to the experiment, etc.” facilitates the reproducibility of the whole process.

Data and workflows deployed in an AI system are two key ingredients in traceability and provenance tracking. In particular, the distinction between *prospective* and *retrospective* provenance is introduced in the literature when dealing with workflows. The prospective provenance models workflows in an abstract and informative way, as templates composed of tasks that can be instantiated, modified, and combined. The retrospective provenance models past workflows, highlighting what task was executed and how data or other artifacts were derived [4]. Herschel et al. [1] claims that “provenance can be seen as metadata that, instead of describing data, describes a production process”. Considering the central role metadata play in tracking provenance, this section discusses some popular models to track provenance. In particular, Garijo et al. [5] have provided a holistic, Linked Data compliant, and ready-to-use solution to document workflow specifications and their executions, which exploits PROV-O [6], P-PLAN [7] and the Open Provenance Model for Workflows (OPMW)[1].

PROV is a metadata model defined as a W3C recommendation. It captures the provenance documenting the entities, agents, actions, and the involved in a production chain, and the relations among them (e.g., attribution and usage). PROV acknowledges the need to represent workflows, also called plans, by including a construct such as *prov:Plan*. However, “it does not elaborate any further on how plans can be described or related to other provenance elements of the execution.” [7].

P-PLAN vocabulary extends PROV-O introducing constructs for plans (*p-plan:Plan* subclass of *prov:Plan*), their steps (*p-plan:Step*) and their input and output variables (*p-plan:Variable*). Still, P-PLAN does not model a full-fledged notion of workflow.

OPMW extends P-PLAN and the “Open Provenance Model (OPM), a legacy provenance model developed by the workflow community that was used as a reference to create PROV” [5]. OPMW distinguishes between workflow specifications, namely *templates*, and their workflow execution traces.

OPMW specifies workflow *templates* as instances of the class *opmw:WorkflowTemplate* (subclass of *p-plan:Plan*); the template processes/actions as *opmw:WorkflowTemplateProcess* (subclass of *p-plan:Step*); the template artifacts, manipulated or produced by processes, as *opmw:WorkflowTemplateArtifact* (subclass of *p-plan:Variable*). Accordingly, the template for the generic n-th step is an instance of *opmw:WorkflowTemplateProcess*. The n-th template steps’ input and output are indicated by the properties *p-plan:hasInputVar* and *p-plan:isOutputVarOf*, and are instances of *opmw:WorkflowTemplateArtifact*, representing any expected file, parameter, and collection of documents considered and manipulated by the template step. The classes *opmw:WorkflowExecutionAccount*, *opmw:WorkflowExecutionProcess* and *opmw:WorkflowExecutionTemplate* represent the execution counterparts of the template instances. The properties *opmw:correspondsToTemplate*, *opmw:correspondsToTemplateProcess*, *opmw:correspondsToTemplateArtifact* bind the execution and the template counterparts. Thus, n-th step is the actual execution of the n-th template step and it is an instance of *opmw:WorkflowExecutionProcess*, which is a specialization of the class *prov:Activity*. The actual execution’s n-th input and output steps are indicated by the PROV properties *prov:used* and *prov:wasGeneratedby*, and are instances of *opmw:workflowExecutionArtifact* which is a particular kind of *prov:Entity*. Albertoni et al. [8] provides examples of the use of the above metadata models when documenting scientific experiments.

Although not specific to AI experiments and systems, the models mentioned above offer some excellent standing and a backbone for describing data, actors, other kinds of entities, and how these might relate in experiments. Such a standing needs to be refined and extended to capture the gist of specific AI experiments. AI-related controlled terminologies might be required, for example, to complements the backbones with the hyper-parameters, tasks and metrics for AI techniques. Adopting a backbone, which is defined according to linked data best practices, offers the ability to combine different models and terminologies as needed, easing the tailoring of such backbone with the required AI-specific and community-governed refinements.

Bibliography

[1](1,2,3) Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? What form? What from? *VLDB Journal*, 26(6):881–906, 2017. doi:10.1007/s00778-017-0486-1.

[[2]] Marçal Mora-Cantallops, Salvador Sánchez-Alonso, Elena García-Barriocanal, and Miguel-Angel Sicilia. Traceability for trustworthy AI: a review of models and tools. *Big Data and Cognitive Computing*, 2021. URL: https://www.mdpi.com/2504-2289/5/2/20, doi:10.3390/bdcc5020020.

[[3]] Yolanda Gil, Ewa Deelman, Mark H. Ellisman, Thomas Fahringer, Geoffrey C. Fox, Dennis Gannon, Carole A. Goble, Miron Livny, Luc Moreau, and Jim Myers. Examining the challenges of scientific workflows. *IEEE Computer*, 40(12):24–32, 2007. URL: https://doi.org/10.1109/MC.2007.421, doi:10.1109/MC.2007.421.

[[4]] Chunhyeok Lim, Shiyong Lu, Artem Chebotko, and Farshad Fotouhi. Prospective and retrospective provenance collection in scientific workflow environments. In *IEEE SCC*, 449–456. IEEE Computer Society, 2010. URL: http://dblp.uni-trier.de/db/conf/IEEEscc/scc2010.html#LimLCF10.

[5](1,2) Daniel Garijo, Yolanda Gil, and Óscar Corcho. Abstract, link, publish, exploit: an end to end framework for workflow sharing. *Future Generation Comp. Syst.*, 75:271–283, 2017. doi:10.1016/j.future.2017.01.008.

[[6]] Deborah McGuinness, Timothy Lebo, and Satya Sahoo. PROV-o: the PROV ontology. W3C Recommendation, W3C, April 2013. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/.

[7](1,2) Daniel Garijo and Yolanda Gil. Augmenting PROV with Plans in P-PLAN: scientific processes as linked data. In *Proceedings of the 2nd International Workshop on Linked Science*, volume 951 of CEUR Workshop Proceedings. 2012. URL: http://oa.upm.es/19478/.

[[8]] Riccardo Albertoni, Monica De Martino, and Alfonso Quarati. Documenting context-based quality assessment of controlled vocabularies. *IEEE Trans. Emerg. Top. Comput.*, 9(1):144–160, 2021. URL: https://doi.org/10.1109/TETC.2018.2865094, doi:10.1109/TETC.2018.2865094.

This entry was written by Riccardo Albertoni.

[[1]] http://www.opmw.org/model/OPMW/

Continuous Performance Monitoring

In brief

Continuous performance monitoring is the activity to track, log and monitor over time the behaviour and the performance of Artificial Intelligence and Machine Learning models. This activity is particularly relevant after in-production deployment in order to detect any performance drifts and outages of the model.

More in Detail

Monitoring the live functioning of a produtionalised ML/AI model or system is an emergent topic that is gaining increasing attention as more and more methods are being deployed in industrial, commercial and public sectors. As any other piece of software, any tool based on AI/ML needs to be maintained over time, for fixing bugs and ensuring quality. ML models and systems require specific strategies that take into account their nature of learning from data.

Idealistically, the behaviour of ML models trained on sample well-curated data is expected to generalise on new, unseen data in the post-deployment phase. Nonetheless, this happens rarely in practice, and a model's performance assessed live is often different from the performance evaluated offline during development. Furthermore, it is well-known that the performance of an AI model or system degrades over time.

Several phenomena have been identified as drivers of this decay. The input data fed into the ML model may contain unexpected patterns not present in the training datasets. Moreover, the characteristics of data may change over time, causing that the relationships at the core of the ML methods do not stand valid any more.

This phenomenon, termed *concept* or *model drift* [1], can lead the model to make wrong predictions. Additionally, if the nature (or distribution) of the input data become vastly different with respect to those used for training, the performance can even drop below acceptance. This phenomenon is known as *covariate shift* [2]. Performance degradation can also result from the impact that the same deployed ML model may have on the decision process that it supports. The ML model may influence other elements involved in the decision or induce an overall change in the phenomenon that is being modelled, which was not taken into account during training.

Overall, after its deployment, an ML method can come across several difficulties and changes that the level of efforts and skills needed in its maintenance could be an order of magnitude higher than that needed in model building.

Given these concerns, several strategies and best practices have been investigated to monitor the behaviour of ML methods after deployment, also in relation to any consequence the methods can have. The first work published in 2015 described the various challenges that ML methods raise after deployment in relation to data dependencies, model complexity, reproducibility, testing, and changes in the external world [3]. After that, several methods have been presented in the literature, focusing specifically on data [4], on the role of humans in ML deployment [5], on testing strategies [6], or the definition of a general framework to track ML methods in their live functioning (e.g., pipelines, datasets, execution configurations, code and human actions) [7].

Overall, the best practices, promoted also from industrial actors [8, 9], include a continuous monitoring of the ML system to assess its quality and “vitality”. Various types of metrics are suggested in this respect, focusing mainly on performance evaluation. The idea is to detect changes in the behaviour and then act via re-training or implementing an active learning approach (when reinforcement learning is adopted), so as to rectify any wrong conduct. It should be noted that model maintenance can be seen as nurturing the model, as it can take advantage of the new knowledge coming from the real-setting scenario, thus it can produce an improvement of the original version released.

Monitoring and maintenance can be performed in a *proactive* or *reactive* fashion. Proactive monitoring works to identify the input samples that deviate significantly from the patterns seen in the training phase and to analyse them more in detail to understand any drifts. The reactive approach entails detecting a wrong output and identifying its causes, so as to understand how the method can be rectified.

The Continuous Delivery [10] and DevOps [11] approaches have been also proposed to better manage the risks of releasing changes to Machine Learning applications and, then, do them in a safe and reliable way.

Bibliography

[[1]] Alexey Tsymbal. The problem of concept drift: definitions and related work. 2004.

[[2]] Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments*. MIT Press, 2012. ISBN 9780262017091.

[[3]] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

[[4]] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, 1723–1726. New York, NY, USA, 2017. Association for Computing Machinery. URL: https://doi.org/10.1145/3035918.3054782, doi:10.1145/3035918.3054782.

[[5]] Ilias Flaounas. Beyond the technical challenges for deploying machine learning solutions in a software company. 2017. URL: https://arxiv.org/abs/1708.02363, doi:10.48550/ARXIV.1708.02363.

[[6]] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D. Sculley. What’s your ml test score? a rubric for ml production systems. In *NIPS 2016 Workshop (2016)*. 2016.

[[7]] Vinay Sridhar, Sriram Subramanian, Dulcardo Arteaga, Swaminathan Sundararaman, Drew Roselli, and Nisha Talagala. Model governance: reducing the anarchy of production ML. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*, 351–358. Boston, MA, jul 2018. USENIX Association. URL: https://www.usenix.org/conference/atc18/presentation/sridhar.

[[8]] Accenture. Model behavior. nothing artificial. 2017.

[[9]] SaS. Machine learning model governance. white paper. 2019.

[[10]] Wolff. *A Practical Guide to Continuous Delivery*. Addison-Wesley, 2017. ISBN.

[[11]] Loukides. *MWhat is DevOps?* O'Reilly Media, 2012. ISBN.

This entry was written by Sara Colantonio.

Respect for Privacy

In Brief

Respect for Privacy is an ethical aspect studied in the TAILOR project for ensuring personal data protection that is at the core of the General Data Protection Regulation (GDPR) [1]. GDPR, in its Article 5, promote *privacy by design* in the form of a certain number of general principles for ensuring *privacy as the default* in the whole chain of data processing for a given task. We outline the challenges and solutions for enforcing privacy by design approaches.

Abstract

When protecting personal data, we are faced to the dilemma of disclosing no sensitive data while learning useful information about a population. One approach for solving this tension between privacy and utility is based on *data encryption* and consists in developping secure computation protocols able to learn models or to compute statistics on encrypted data. A lot of scientific literature [2] [3] [4] [5] has been exploring this security-based approach depending on the target computational task. Another approach consists in conducting data analysis tasks on datasets made anonymous by the application of some ./T3.5/L1.privacy_mechanisms. according to some Privacy Models. Anonymization must no be reduced to Pseudonymization (see also Re-identification Attack), which is defined in GDPR Article 4 as “the processing of personal data in such a way that the data cannot be attributed to a specific data subject without the use of additional information.” Anonymization, as defined in GDPR (see Recital 26), refers to a process that removes any possibility of identifying a person even with additional information. In the resulting anonymized data, the connection should be completely lost between data and the individuals. Based on such a definition, anonymization is very difficult to model formally and to verify algorithmically. We will briefly survey the main privacy models and their properties, as well as the main privacy mechanisms which can be applied for enforcing the corresponding privacy properties or for providing strong guarantees of robustness to attacks.

Motivation and Background

Publishing datasets plays an essential role in open data research and in promoting transparency of government agencies. Unfortunately, the process of data publication can be highly risky as it may disclose individuals’ sensitive information. Hence, an essential step before publishing datasets is to remove any uniquely identifiable information from them. This is called Pseudonymization and consists in masking or replacing by pseudonyms values of properties that directly identify persons such as their name, address, postcode, telephone number, photograph or image, or some other unique personal characteristic.

Pseudonymization is not sufficient however for preserving the privacy of users. Adversaries can re-identify individuals in datasets based on common attributes called quasi-identifiers or may have prior knowledge about the users. Such side information enables them to reveal sensitive information that can cause physical, financial, and reputational harms to people.

Therefore, it is crucial to assess carefully *privacy risks* before the publication of datasets. Detection of privacy breaches should come with explanations that can then be used to guide the choice of the appropriate anonymization mechanisms to mitigate the detected privacy risks. *Anonymization* should provide provable guarantees for privacy properties induced by some Privacy Models. Differential Privacy and k-anonymity are the two main privacy models for which ./T3.5/L1.privacy_mechanisms have been designed. They enjoy different properties based on the type of perturbations or transfomations applied on the data to anonymize.

The strength of Pseudonymization and anonymization techniques can be assessed by their robustness to privacy attacks that aim at re-identifying individuals in datasets based on common attributes called quasi-identifiers or on prior knowledge.

Guidelines

EDBP has published several guidelines. The EDPB Guidelines on Data Protection Impact Assessment focus on determining whether a processing operation is likely to result in a high risk to the data subject or not. It provides guidance on how to assess data protection risks and how to carry out a data protection risk assessment.

Data minimisation is a strong recommendation to limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose, and to retain the data only for as long as is necessary to fulfil that purpose.

The most authoritative guideline on data protection “by design and by default” outlines the data subject’s rights and freedoms and the data protection principles that are illustrated through examples of practical cases. It emphasizes the obligation for controllers to stay up to date on technological advances on handling data protection risks, and to implement and update the measures and safeguards taking into account the evolving technological landscape.

Software Frameworks Supporting Dimension

There are some practical tools that help in enanching respect for privacy and awareness, in particular definining and mitigating potential privacy risks. This is compliant with the Data Protection Impact Assessment introduced in the GDPR.

Risks can be identified and addressed at an early stage by analyzing how the proposed uses of personal information and technology will work in practice. We should identify the privacy and related risks, evaluate the privacy solutions and integrate them into the project plan.

Currently, a lot of frameworks have implemented to manage this task. University of British Colombia provides a tool for determining a project’s privacy and security risk classification. TrustArc offers a consulting service for analyzing personally identifiable information, looking at risk factors and assisting in the development of policies and training programs. Information Commissioner’s Office provides a handy step by step guide through the process of deciding whether to share personal data[1][2]. Also the US Department of Homeland Security implemented such decision tool.

A (non-exhaustive) list of more practical tools includes:

- Amnesia, a tool for anonymize tabular data relying on k-anonymity paradigm;
- AXR, a tool that incorporates different Privacy Models;
- Scikit-mobility, a Python library for mobility analysis that includes the computation of privacy risks in such setting, based on the work presented in [6] [7].

Taxonomic Organisation of Terms

The *Respect for Privacy* dimension mainly regards the Data Protection. The Assessment of Privacy Risks can be performed, and two diffent strategies are available two protect the data privacy. The first one regards the application of Anonymization Mechanisms, such as Pseudonymization, k-anonymity, or Differential Privacy. The second strategy is Data Encryption, which is strictly related with the Security Dimension. In Fig. 18, one can find the taxonomy proposed here. In blue, there are highlighted the possible attacks related to the various strategies, i.e., Re-identification Attack, ./T3.5/L2.membership, and Security Attacks.

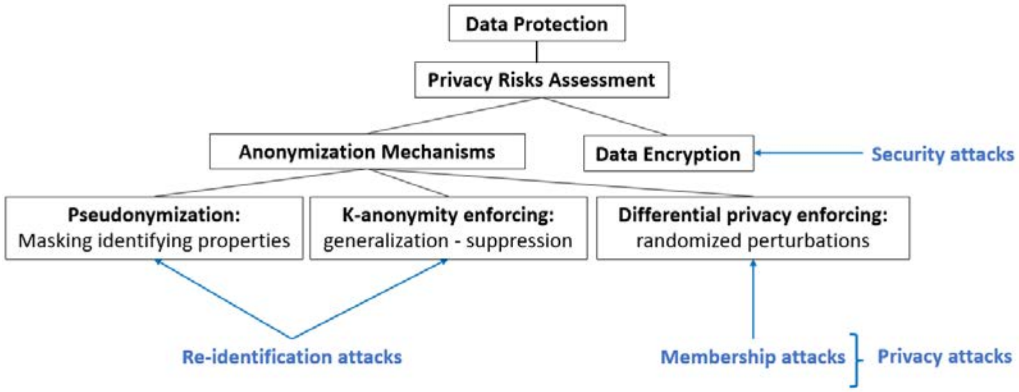


Fig. 18 A possible taxonomy about solutions to the Respect for Privacy dimension.

Main Keywords (TBA)

- Data Anonymization (and Pseudonymization): A data subject is considered anonymous if it is *reasonably* hard to attribute his personal data to him/her.
- Pseudonymization: **Pseudonymisation** aims to substitute one or more identifiers that link(s) the identity of an individual to its data with a surrogate value, called **pseudonym** or **token**.
- Privacy Models: There are essentially two families of models, based on different goals and mechanisms: anonymity by randomization (where the most recent paradigm is Differential Privacy) and anonymity by indistinguishability (whose most famous example is k-anonymity).
- Differential Privacy: **Differential privacy** implies that adding or deleting a single record does not significantly affect the result of any analysis.
- \epsilon-Differential Privacy: **\(\epsilon\)-Differential Privacy** is the simpler form of Differential Privacy, where \(\epsilon\) represents the level of privacy guarantee.
- (\epsilon,\delta)-Differential Privacy: A relaxed version of Differential Privacy, named **(\(\epsilon\),\(\delta\))-Differential Privacy**, allows a little privacy loss (\(\delta\)) due to a variation in the output distribution for the privacy mechanism.
- Achieving Differential Privacy: Differential privacy guarantees can be provided by perturbation mechanisms aim at randomizing the output distributions of functions in order to provide privacy guarantees.
- k-anonymity: **k-anonimity** (and the whole family of **anonymity by indistinguishability** models) is based on comparison among individuals present in data, and it aims to make each individual so similar as to be indistinguishable from at least *k-1* others.
- Attacks on anonymization schemes: There are a variety of attacks that involve data privacy. Some of them are very context-specific (for example, there exists attacks on partition-based algorithms, such as deFinetti Attack or Minimality Attack), while other are more general.
- Re-identification Attack: **Re-identification attack** aims to link a certain set of data related to an individual in a dataset (which does not contain direct identifiers) to a real identity, relying on additional information.

Bibliography

[[1]] European Parliament & Council. General Data Protection Regulation. 2016. L119, 4/5/2016, p. 1–88.

[[2]] C. Castelluccia, A. C.-F. Chan, E. Mykletun, and G. Tsudik. Efficient and Provably Secure Aggregation of Encrypted Data in Wireless Sensor Networks. *ACM Transactions on Sensor Networks (TOSN)*, 5(3):20:1–20:36, 2009.

[[3]] Q.-C. To, B. Nguyen, and P. Pucheral. Private and Scalable Execution of SQL Aggregates on a Secure Decentralized Architecture. *ACM Transactions on Database Systems (TODS)*, 41(3):16:1–16:43, 2016.

[[4]]

J. Mirval, L. Bouganim, and I. Sandu Popa. Practical Fully-Decentralized Secure Aggregation for Personal Data Management Systems. In *International Conference on Scientific and Statistical Database Management (SSDBM)*, 259–264. 2021.

[[5]] R. Ciucanu, M. Giraud, P. Lafourcade, and L. Ye. Secure Grouping and Aggregation with MapReduce. In *International Joint Conference on e-Business and Telecommunications (ICETE) – Volume 2: International Conference on Security and Cryptography (SECRYPT)*, 514–521. 2018.

[[6]] Luca Pappalardo, Filippo Simini, Gianni Barlacchi, and Roberto Pellungrini. Scikit-mobility: a python library for the analysis, generation and risk assessment of mobility data. 2019. https://arxiv.org/abs/1907.07062.

[[7]] Francesca Pratesi, Anna Monreale, Roberto Trasarti, Fosca Giannotti, Dino Pedreschi, and Tadashi Yanagihara. PRUDEnce: a system for assessing privacy risk vs utility in data sharing ecosystems. *Transaction Data Privacy*, 11(2):139–167, 2018.

This entry was written by Marie-Christine Rousset, Tristan Allard, and Francesca Pratesi.

[[1]] https://ico.org.uk/for-organisations/sme-web-hub/checklists/data-protection-self-assessment/data-sharing-and-subject-access-checklist/

[[2]] https://ico.org.uk/for-organisations/guide-to-data-protection/ico-codes-of-practice/data-sharing-a-code-of-practice/

Data Anonymization (and Pseudonymization)

In Brief

A data subject is considered anonymous if it is *reasonably* hard to attribute his personal data to him/her.

More in Detail

A data subject is considered anonymous if it is *reasonably* hard to attribute his personal data to him/her. What “reasonably” actually means depends both on the context and on the requirements given by data respondents. Both the identity of a subject and other information related to him/her are considered in the context of anonymity, for example sensitive information regarding health, religion, political tendencies and, more in general, any kind of information that can somehow distinguish the subject from others. The European General Data Protection Regulation (GDPR [1]) defines anonymous data as

> “[…] information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable”.

Thus, data to have this property has to be deprived of all distinctive elements of a person, i.e., those elements that permit to identify both directly or indirectly that person in the data. Since anonymous data does not enable re-identification of data subjects, even with the use of additional information, this type of data is not subject to the privacy regulations since it is not considered personal data.

Note that this is process is absolutely different from removing the direct identifiers only (e.g., name, surname, social security number). This process is called *de-identification*, and it can be subjected to privacy leaks, such as re-identification. The process of substitute a set of direct identifiers with a surrogate value, or psudonym, is called Pseudonymization. It suffers from the same weaknesses of de-identification. Nevertheless, the GDPR strongly advocates the application of Pseudonymization:

> “The application of pseudonymisation to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations” (Recital 28).

> “In order to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures […]. Such measures could consist, inter alia, of […] pseudonymising personal data as soon as possible […]” (Recital 78).

> “The further processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes is to be carried out […] provided that appropriate safeguards exist (such as, for instance, pseudonymisation of the data)” (Recital 156).

This entry was readapted from *Comandé et al. Elgar Encyclopedia of Law and Data Science. Edward Elgar Publishing (2022) ISBN: 978 1 83910 458 9* by Francesca Pratesi, Roberto Pellungrini, and Anna Monreale.

Pseudonymization

In Brief

Pseudonymisation aims to substitute one or more identifiers that link(s) the identity of an individual to its data with a surrogate value, called **pseudonym** or **token**.

More in detail

To preserve a subject’s privacy, one of the most basic methodology is to de-couple the identity of said subject from its data. This is process is called pseudonymisation. The typical practical approach to achieve pseudonymity is to detect which attributes in the data may reveal the subject’s identity, called *personal identifiers*, and substitute them with some other value.

However, re-identification may be needed in certain cases (for example, to contact data subject for further questions), therefore personal identifiers are often maintained for re-associating subject and identity. This association should be secured and inaccessible to anybody having access tho the pseudonymised data, so that protection is guaranteed.

Following the description in the Article 4(5) of the European General Data Protection Regulation (GDPR) [1]:

> “the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person”.

This definition indicates that the additional information needed to actually link a subject’s identity to it’s data should be the focus of pseudonymisation techniques. Indeed, pseudonymization reduces the risk of publishing data due to a direct re-identification (See Re-identification Attack).

Recital 28 of the GDPR, states that “explicit introduction of pseudonymisation in this Regulation is not intended to preclude any other measures of data protection”. Article 6(4) of the GDPR also reports that pseudonymisation could be an “appropriate safeguards” and that data controller should operate “pseudonymising personal data as soon as possible” (Recital 78 GDPR) and implementing “appropriate technical and organisational measures, such as pseudonymisation” (Article 25(1) GDPR) both at the time of the determination of the means for processing and at the time of the processing itself.

The pseudonym must be distinguishable and irreversible in the absence of additional information. This means that it should not be possible to reconstruct the original value by just considering the pseudonym, i.e., there does not exist any function that computes the original value with the pseudonym as input. The correspondence between original value and pseudonym must be stored in a separate location and must be secured against data breaches. Surrogate values need also to be managed after the generation, either internally or externally. In the latter case, the institution who owns data outsources this service to a qualified (and trusted) third party.

Pseudonymisation techniques

There are several techniques that perform pseudonymisation. They can be generally summarized in three main categories:

1. **Cryptographic with secret key**: these techniques use mathematical mechanism to alter the original value through the application of a secret *key*. This key is at the core of the mechanism: with it, the pseudonymisation can be reversed, so the key has to be secured at all times.
2. **Hash-based**: these techniques use a function that, given an identifier (composed by one or more attributes) with arbitrary length returns a value of fixed size (e.g., size 256 bits, which correspond to 32 characters), being called hash value or message digest. The hash function is usually a deterministic function and must be irreversible, i.e., for any input of the function it is infeasible to compute the inverse function from the output. Functions typically used for hashing are *SHA-2* [2] and *SHA-3* [3], for example the *SHA3-512* which has output values of length 512 bits.
3. **Keyed-hash based**: a combination of the previous techniques where the hash function requires a key, called *salt* [4], to compute its output. This is generally considered a more robust approach that simple hashing. Varying the key, the same data subject's identifier can be translated in several different pseudonyms. In cryptography literature, these are referred to as *message authentication codes* [5]. This family of techniques is more robust against some brute-force attacks, especially if the salt is changed sufficiently often.
4. **Keyed-hash function with deletion**: equal to the previous one, but after the generation of the pseudonym, the correspondence table is deleted, i.e., we cannot associate again pseudonyms to personal identifiers.
5. **Tokenization**: the idea of tokenization is to substitute the subjects' identifiers with a token generated with some cryptographic methods. However, tokenization is a non-mathematical approach: data is replaced but the type or length is not altered. Typically knowledge of a token has no usefulness for a third party. Another difference is that tokenization is fast and can be done with few computational resources.[6]

Bibliography

[[1]] European Parliament & Council. General Data Protection Regulation. 2016. L119, 4/5/2016, p. 1–88.

[[2]] Information Technology Laboratory, National Institute of Standards and Technology. Secure hash standard (SHS). URL: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf (visited on 2022-05-02).

[[3]] Information Technology Laboratory, National Institute of Standards and Technology. SHA-3 standard: permutation-based hash and extendable-output functions. URL: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.202.pdf (visited on 2022-05-02).

[[4]] Information Technology Laboratory, National Institute of Standards and Technology. The keyed-hash message authentication code (hmac). URL: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.198-1.pdf (visited on 2022-05-02).

[[5]] Alfred J. Menezes, Scott A. Vanstone, and Paul C. van Oorschot. *Handbook of Applied Cryptography*. CRC Press, 1996. URL: https://cacr.uwaterloo.ca/hac/ (visited on 2022-05-02).

[[6]] Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques. URL: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (visited on 2022-05-02).

This entry was readapted from *Comandé et al. Elgar Encyclopedia of Law and Data Science. Edward Elgar Publishing (2022) ISBN: 978 1 83910 458 9* by Francesca Pratesi, Roberto Pellungrini, and Anna Monreale.

Privacy Models

There are essentially two families of models, based on different goals and mechanisms: anonymity by randomization (where the most recent paradigm is Differential Privacy) and anonymity by indistinguishability (whose most famous example is k-anonymity).

This entry was written by Francesca Pratesi.

Differential Privacy

In brief

Differential privacy implies that adding or deleting a single record does not significantly affect the result of any analysis.

More in detail

The Family of Differential Privacy Models

Differential privacy is a prominent family of privacy-preserving data publishing models (see Privacy Models). It comprehends privacy as the ability to set a limit on the impact of any single individual on the outputs of the function that produces the information to publish (computation, e.g., of a set of statistics, of a machine learning model, of generated synthetic data). In other words, a differentially private function promises to each individual that its outputs will be more or less the same whether the individual's data is input by the function or not. Differential privacy models all share this common intuitive goal but they differ in the way they formalize it - for example, on the quantification of the impact of an individual or on the tolerance to possible failures of the guarantees (though improbable). They usually exhibit properties that have been identified as key requirements to privacy models. <!–(see ./L2.privmod_properties for details).

Achieving Differential Privacy

Designing a function that satisfies differential privacy often boils down to carefully combining basic perturbation mechanisms (such as, e.g., the Laplace mechanism) and to demonstrating formally either that data is only accessed through a differentially private function (leveraging the safety under post-processing and the self-composability properties), or that the output distribution of the function complies with the targeted differential privacy model (through, e.g., randomness alignments). We refer the interested reader to Achieving Differential Privacy for more information.

An expanding universe

The seminal differential privacy models were proposed in the mid-2000's and include \(\epsilon\)-differential privacy (see \(\epsilon\)-Differential Privacy) or \((\epsilon, \delta)\)-differential privacy (see (\(\epsilon\),\(\delta\))-Differential Privacy). The number of differential privacy models has grown fastly over the years (more than 200 extensions or variants have been reported in a 2020 survey paper). Differential privacy is often considered in the academia as a *de facto* standard for privacy-preserving data publishing and has earned the original authors the prestigious Gödel Prize in 2017. Famous organizations (e.g., the US Census Bureau) and companies (e.g., Google, Apple, LinkedIn, Microsoft) have launched ambitious real-life applications of differential privacy.

Bibliography

The seminal differential privacy models were introduced in [1], [2], and [3]. Differential privacy is thoroughly introduced in [4] and numerous variants and extensions are surveyed in [5]. The book [6] surveys differential privacy techniques related to database queries.

[[1]] Cynthia Dwork. Differential privacy. In *ICALP*. 2006.

[[2]] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*. 2006.

[[3]] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: privacy via distributed noise generation. In *EUROCRYPT*. 2006.

[[4]]

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, 2014.

[[5]] Damien Desfontaines and Balázs Pejó. Sok: differential privacies. *Proceedings on Privacy Enhancing Technologies*, 2020:288 – 313, 2020.

[[6]] Joseph P. Near and Xi He. Differential privacy for databases. *Foundations and Trends in Databases*, 11:109–225, 2021.

This entry was written by Tristan Allard.

\(\epsilon\)-Differential Privacy

Synonyms: \(\epsilon\)-indistinguishability.

In brief

\(\epsilon\)-Differential Privacy is the simpler form of Differential Privacy, where \(\epsilon\) represents the level of privacy guarantee.

More in Detail

The Statistical Databases Context

The seminal \(\epsilon\)-differential privacy model lays down the basic notions related to differential privacy by formalizing the intuitive requirement that *the possible impact of any single individual on the output of a differentially private function must be limited* (see Differential Privacy). The initial \(\epsilon\)-differential privacy model focuses on the context of statistical databases: the private dataset is a table \(\mathcal{D}\) in which each individual contributes at most one record, the system answers interactively to a sequence of statistical queries over \(\mathcal{D}\) (e.g., a sequence of queries containing counts, sums, averages, *etc*), and the \(\epsilon\)-differential privacy model aims at limiting the information leakage about the private dataset.

Formalizing Differential Privacy

In a nutshell, the \(\epsilon\)-differential privacy model requires that the presence/absence of any *possible* individual does not shift any output probability by more than a factor of \(e^\epsilon\). More precisely, a *random function* \(\mathtt{f}\) with range \(\mathcal{O}\) satisfies \(\epsilon\)-differential privacy if and only if for all possible pairs of datasets (\(\mathcal{D}\), \(\mathcal{D}'\)) such that \(\mathcal{D}'\) is \(\mathcal{D}\) with *one record more or one record less*, and for all \(\mathcal{S} \subseteq \mathcal{O}\), then it holds that: \(\mathtt{Pr} [\mathtt{f} (\mathcal{D}) \in \mathcal{S}] \leq e^\mathbf{\epsilon} \times \mathtt{Pr} [\mathtt{f} (\mathcal{D}') \in \mathcal{S}]\) where \(\epsilon>0\) is the privacy parameter.

Let us comment the above definition. First, the function \(\mathtt{f}\) can be any arbitrary function, including the usual statistical functions (e.g., counts, sums) but not restricted to them. Second, the pairs of datasets whose output distributions must not differ too much are taken from the full space of the possible datasets; they are not derived from the actual private dataset. Third, the impact of an individual is defined based on the presence (or absence) of his/her record in (or from) any possible dataset. Pairs of datasets that differ on the presence/absence of a single record are called *neighboring datasets*. Note that variants might exist (e.g., by considering that neighboring datasets are datasets that differ on the value of a single row). Fourth, the value of \(\epsilon\) sets the tolerance of the model to the possible impacts of individuals on the output of \(\mathtt{f}\): the lower the \(\epsilon\) the more stringent the requirement. Common values range from \(\epsilon=0.01\) to \(\epsilon=10\).

Please see Achieving Differential Privacy for a synthesis of **how** common functions can be adapted in order to satisfy \(\epsilon\)-differential privacy.

Self-Composability and Safety Under Post-Processing

The \(\epsilon\)-differential privacy model is self-composable as follows. The *parallel composition* of two functions, respectively satisfying \(\epsilon_1\)-differential privacy and \(\epsilon_2\)-differential privacy, satisfies \(\max (\epsilon_1, \epsilon_2)\)-differential privacy. Their *sequential composition* satisfies \((\epsilon_1 + \epsilon_2)\)-differential privacy. The \(\epsilon\)-differential privacy model is as well convex and safe under post-processing.

Bibliography

The ϵ-differential privacy model was introduced in [1] and the ϵ-indistinguishability model in [2].

[[1]] Cynthia Dwork. Differential privacy. In *ICALP*. 2006.

[[2]] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*. 2006.

This entry was written by Tristan Allard.

(ϵ,δ)-Differential Privacy

In brief

A relaxed version of Differential Privacy, named **(ϵ,δ)-Differential Privacy**, allows a little privacy loss (δ) due to a variation in the output distribution for the privacy mechanism.

More in Detail

Relaxing ϵ-Differential Privacy
The (ϵ,δ)-differential privacy model is a common relaxation of ϵ-differential privacy. Under the ϵ-differential privacy model, the probabilities that the function $\mathtt{f}$ outputs the same output when computed over neighboring datasets are allowed to diverge up to an e^ϵ factor. The (ϵ,δ)-differential privacy model additionally tolerates the two probabilities to diverge by a small additional quantity, denoted δ.

This leads to revisiting the formal definition of ϵ-differential privacy as follows. A *random function* $\mathtt{f}$ with range $\mathcal{O}$ satisfies (ϵ, δ)-differential privacy if and only if for all possible pairs of datasets ($\mathcal{D}$, $\mathcal{D}'$) such that $\mathcal{D}'$ is $\mathcal{D}$ with one record more or one record less, and for all $\mathcal{S} \subseteq \mathcal{O}$, then it holds that: $\mathtt{Pr} [\mathtt{f} (\mathcal{D}) \in \mathcal{S}] \leq e^\mathbf{\epsilon} \times \mathtt{Pr} [\mathtt{f} (\mathcal{D}') \in \mathcal{S}] + \delta$ where $\epsilon>0$ and $\delta \geq 0$ are the privacy parameters. When $\delta>0$, (ϵ,δ)-differential privacy is also called *approximate differential privacy*.

Bibliography

The (ϵ,δ)-differential privacy model is introduced in [3] and thoroughly studied in [2].

[[1]] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: privacy via distributed noise generation. In *EUROCRYPT*. 2006.

[[2]] Sebastian Meiser. Approximate and probabilistic differential privacy definitions. *IACR Cryptol. ePrint Arch.*, 2018:277, 2018.

This entry was written by Tristan Allard.

Achieving Differential Privacy

In Brief

Differential privacy guarantees can be provided by perturbation mechanisms aim at randomizing the output distributions of functions in order to provide privacy guarantees.

More in Detail

Perturbation mechanisms aim at randomizing the output distributions of functions in order to provide privacy guarantees. We focus on the major mechanisms able to provide differential privacy guarantees (see Differential Privacy) and concentrate on the major mechanisms. *The Laplace Mechanism* is dedicated to perturb functions that output real values, allowing them to satisfy \(\epsilon\)-Differential Privacy. The Exponential Mechanism is a generalization of the Laplace mechanism and also allows to satisfy \(\epsilon\)-differential privacy. The éGaussian Mechanism* is a variant of the Laplace mechanism, applying to functions that output real values as well, but allowing them to satisfy the \((\epsilon,\delta)\)-differential privacy relaxation. These three mechanisms are appropriate for perturbing centralized functions that input a full dataset (e.g., a sum query). They are often used as basic building blocks, combined or not, for perturbing elaborate functions. Finally, Randomized Response Mechanisms input and output one single row at a time (represented as a vector of bits): they are local mechanisms and can be applied prior to the data collection.

History

The earliest known randomized response mechanism has been proposed in the 1960's (decades before differential privacy) by Warner, a sociologist who wanted to improve the reliability of responses to sensitive questions by letting the interviewee perturb his/her answer in a controlled manner. Thanks to the simplicity of their implementation and to the differential privacy guarantees that they provide, they generate a renewed interest both from the academia and from the industry.

The Laplace mechanism (resp. the Gaussian mechanism) has been proposed jointly with the seminal \(\epsilon\)-differential privacy model (resp. the \((\epsilon, \delta)\)-differential privacy model). Its generalization as the Exponential mechanism was proposed the following year.

Bibliography

The early randomized response mechanism was proposed in [3], the Laplace mechanism in [2] and the Exponential mechanism in [4]. The Gaussian mechanism was shown to satisfy \((\epsilon, \delta)\)-differential privacy in [3]. Finally, [5] provides an overview and an evaluation of various randomized response mechanisms proposed before 2021.

[[1]] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*. 2006.

[[2]] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: privacy via distributed noise generation. In *EUROCRYPT*. 2006.

[[3]] Stanley L. Warner. Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60 309:63–6, 1965.

[[4]] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103, 2007.

[[5]] Graham Cormode, Samuel Maddock, and Carsten Maple. Frequency estimation under local differential privacy. *Proc. VLDB Endow.*, 14(11):2046–2058, 2021.

This entry was written by Tristan Allard.

k-anonymity

In Brief

k-anonimity (and the whole family of **anonymity by indistinguishability** models) is based on comparison among individuals present in data, and it aims to make each individual so similar as to be indistinguishable from at least *k-1* others.

More in Detail

One of the most common models for anonymization is *k*-anonymity, where a subject is considered anonymous if an adversary cannot achieve identification of her within a set of other *k-1* subjects. This set of *k* individuals is called the *anonymity set*. Identification, in this context, means the ability to distinguish the subject from other individuals with whom his/her data are grouped. This definition of anonymity is implicitly quantifiable as it depends from the size of the anonymity set. Clearly, a bigger anonymity set implies a better guarantee regarding indistinguishability. For example, the anonymity of a subject with an anonymity set of size 30 is better than the one that can be achieved with an anonymity set of size 5.

The k-anonymity framework was originally applied on relational tables [1]. The basic assumption is that attributes are partitioned in *quasi-identifiers* and *sensitive attributes* [2]. The quasi-identifiers are attributes that can be linked to external information to re-identify the individual to whom the information refers (so-called Re-identification Attack). Usually, they are publicly known or easy to obtain with a superficial knowledge of the subject: for example, age, zip-code, and gender are classic quasi identifiers. The sensitive attributes, instead, represent the information to be protected. Indeed, a dataset satisfies the property of *k*-anonymity if each released record has at least *k−1* other records also visible in the release whose values are indistinct over the quasi-identifiers.

In *k*-anonymity techniques, strategies such as *generalization* and *suppression* are usually applied to reduce the granularity of representation of quasi-identifiers. It is clear that these methods guarantee privacy but also reduce the accuracy of applications on the transformed data. Indeed, one of the main challenge of *k*-anonymity is to find the minimum level of changes (in terms of generalization or suppression) that allows us to guarantee high privacy and good data precision. In Table 2, one can find an example of a privacy mitigation through generalization that allows to reach *3*-anonymity, i.e., *k*-anonymity with *3* as the minimum size of each anonymity set, starting from the situation depicted in Table 1.

Pseudo-id	Gender	Date of Birth	ZIP code	Disease
1000	F	5 June 1975	02108	Lyme Disease
1001	F	15 May 1973	01970	Hypertension
1002	F	3 December 1977	02657	Vertigo
1003	M	5 September 1941	10238	Stroke
1004	M	15 June 1947	10042	Stroke
1005	M	25 April 1946	10133	Stroke
1006	F	25 December 1942	10053	Diabetes
1007	F	5 October 1949	10053	Osteoporosis
1008	F	6 July 1946	10053	Arthritis

Table 1 A potential extract from a medical dataset. Gender, date of birth, and ZIP code are the quasi-identifiers, while the Disease is the sensitive attribute. An adversary knowing that Alice was born in 1974 and lives in Boston, MA, who gains access to this dataset, will discover that Alice is the patient number 1000 and that she has Lyme Disease.

Pseudo-id	Gender	Date of Birth	ZIP code	Disease
1000	F	[1970-1979]	0****	Lyme Disease
1001	F	[1970-1979]	0****	Hypertension
1002	F	[1970-1979]	0****	Vertigo
1003	M	[1940-1949]	10***	Stroke
1004	M	[1940-1949]	10***	Stroke
1005	M	[1940-1949]	10***	Stroke
1006	F	[1940-1949]	10053	Diabetes
1007	F	[1940-1949]	10053	Osteoporosis
1008	F	[1940-1949]	10053	Arthritis

Table 2 The *3*-anonymous version of the dataset shown in Table 1. The gender attribute is maintained as is, while the date of birth was replaced by an interval of years and the precision of the ZIP code was reduced. An adversary knowing the same information described in Table 1 cannot be sure if Alice is the patient number 1000, 1001 or 1002. Indeed, now each patient is included in an anonymity set of at least *3* individuals.

Weakness of *k*-anonymity model

Unfortunately, the *k*-anonymity framework can be vulnerable in some cases. In particular, it is not safe against *homogeneity attack* and *background knowledge attack*. The homogeneity attack exploits a possible lack of variety in the sensitive attributes. Indeed, an adversary can infer the value of the sensitive attributes when a *k*-anonymous dataset contains a group of *k* entries with the same value for the sensitive attributes. As an example, suppose that the attacker is searching for Bob in the dataset presented in Table 2: the adversary knows that Bob was born in July 1946. Even if the attacker is not able to discriminate Bob between patients 1003, 1004, and 1005, the disease associated with all these three patients is a stroke; thus, the adversary cannot re-identify Bob, but he/she can still infer that Bob suffers from heart disease. In a background knowledge attack, instead, an attacker knows information useful to associate some quasi-identifiers with some sensitive attributes: as an example, remaining in the medical domain, it is common knowledge that certain diseases are more frequent in a specific gender.

Bibliography

[[1]] Latanya Sweeney. Uniqueness of simple demographics in the U.S. population. 2000. LIDAP-WP4.

[[2]] Latanya Sweeney. K-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, oct 2002. URL: https://doi.org/10.1142/S0218488502001648, doi:10.1142/S0218488502001648.

This entry was readapted from *Comandé et al. Elgar Encyclopedia of Law and Data Science. Edward Elgar Publishing (2022) ISBN: 978 1 83910 458 9* by Francesca Pratesi, Roberto Pellungrini, and Anna Monreale.

Attacks on anonymization schemes

There are a variety of attacks that involve data privacy. Some of them are very context-specific (for example, there exists attacks on partition-based algorithms, such as deFinetti Attack or Minimality Attack), while other are more general. For example, we can have the Re-identification Attack, where the aim is to link a identity in the data with a real identity; *Membership Attack*, where the goal is to understand whether or not a particular individual is present in the considered dataset; *Reconstruction Attack*, where one wants to reconstruct (even partially) a private dataset from public aggregate information.

This entry was written by Francesca Pratesi.

Re-identification Attack

Synonyms: Linking attack, Attack on pseudonymised data

In Brief

Re-identification attack aims to link a certain set of data related to an individual in a dataset (which does not contain direct identifiers) to a real identity, relying on additional information.

More in Detail

A first, basic, and simple way to preserve a subject's privacy is to de-couple the identity of said subject from its data. This is process is called Pseudonymization. The typical practical approach to achieve pseudonymity is to detect which attributes in the data may reveal the subject's identity, called *personal identifiers*, and substitute them with some other value. While this process can help in reducing the risk of a direct re-identification of the data subjects based on the published data, re-identification can still be possible in certain cases.

Indeed, additional information (usually called *background information*) can be used by a malicious third party (often called *adversary* or *attacker*) to link a subject's identity to its data. Additional information makes the difference between *pseudonymised* and *anonymised* data. Anonymous data are deprived of all distinctive elements of the person, i.e., those elements that permit to identify both directly or indirectly that person in the data. Anonymous data cannot be re-identified by definition, even when using additional information, therefore this type of data is not subject to the current privacy regulations (e.g., the GDPR [1]).

This key difference can be better understood with the famous real life example of the attack on the privacy of the Governor of Massachussetts. In 1996, *William Floyd Weld*, then Governor of Massachusetts, lost his consciousness during a public event. Rushed to the nearby Deaconess Waltham Hospital, he was officially diagnosed with influenza and consequently discharged the following day [3]. Some time later, professor Latanya Sweeney, a graduate computer science student at MIT at the time, successfully reconstructed what had happened to the governor and inferred his diagnosis linking two different data sources: a publicly available voter rolls dataset and a hospital dataset without patients' names, thus considered anonymous [4]. The voter rolls dataset contained the name, address, ZIP code, birth date, sex and other attributes of every voter in the city of Cambridge (Middlesex County). The hospital dataset was issued to researchers by the Massachusetts Group Insurance Commission and contained diagnosis of patients along with some demographic information. Since the identity of the different patients was not present in the hospital data, the information published there was considered harmless; indeed, this is almost true if thery are considered by themselves. But Sweeney knew that the governor was admitted to the hospital, so she also knew he was present in the data. Therefore, she (in a complete legittimate way) gained access to both datasets and she intersected the demographic information in the two dataset, discoverying some important facts: *six individuals in the hospital dataset shared the Governor's birth date; only three of these were men; but only one of these men lived in the Governor's own ZIP code*.

Surname	Name	Date of birth	Sex	Address	ZIP code	Last vote
…	…	…	…	…	…	…
Weld	William Floyd	31 July 1945	M	75, Essex St	02139	22 May 1998
Welsh	Alice	4 July 1952	F	150, Main St	02139	22 May 1998
Weltcher	Bob	13 July 1947	M	148, Gold Rd	02138	22 May 1998
…	…	…	…	…	…	…

Table 3 Cambridge Voter Roll Dataset: this table represents an extract of the voter dataset.

Id	Sex	Date of birth	ZIP code	Visit	Diagnosis
…	…	…	…	…	…
1	M	31 July 1945	02138	9 May 1996	Diabetes
2	M	31 July 1945	02139	18 May 1996	Stroke
3	F	31 July 1945	02138	18 May 1996	Osteoporosis
4	F	31 July 1945	02139	23 May 1996	Stroke
5	M	4 July 1945	02138	5 June 1996	Diabetes
6	M	13 June 1945	02139	9 June 1996	Arthritis
7	F	5 April 1945	02139	4 July 1996	Hypertension
…	…	…	…	…	…

Table 4 Hospital Dataset: this table represents an extract of the medical dataset. Note that this table does not contain any direct identifiers, such as surnames or social security numbers.

In Table 3 and Table 4 we can see a simplified version of Sweeney's attack. Looking at the tables singularly, no *sensitive information* (i.e., the diagnosis) about the Governor William Floyd Weld can be inferred. However, from Table 3 we gain access to the date of birth and ZIP code of Governor. Then, we can search for the persons born on July 31, 1945 in Table 4, finding patients number 1, 2, 3 and 4. However, ids 3 and 4 correspond to women, so we should consider only individuals 1 and 2. Finally, we can look at the ZIP code in Table 4: the patient number 1 lives in a different area, so we only have one patient (the number 2) that can be the Governor. In brief, we can see that, matching the information colored in blue from the two tables, there is no other possibility for Governor Weld but to be a patient suffering from a stroke. This was a clear breach of the privacy of the Governor, as the public statement about his health differed from the actual cause of hospitalization.

Sweeney conducted similar attacks in a more structured and generalised experiment, finding that 87% of the United States population was uniquely or nearly uniquely identified by the combination of ZIP code, gender, and date of birth [2]. This leaded Sweeney to theorize the k-anonymity principle, and call the attributes used for the re-identification process *quasi-identifier*.

Bibliography

[[1]] European Parliament & Council. General Data Protection Regulation. 2016. L119, 4/5/2016, p. 1–88.

[[2]] Latanya Sweeney. K-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, oct 2002. URL: https://doi.org/10.1142/S0218488502001648, doi:10.1142/S0218488502001648.

[[3]] Daniel C. Barth-Jones. The 're-identification' of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. 2012. doi:http://dx.doi.org/10.2139/ssrn.2076397.

[[4]] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997. PMID: 11066504. URL: https://doi.org/10.1111/j.1748-720X.1997.tb01885.x, arXiv:https://doi.org/10.1111/j.1748-720X.1997.tb01885.x, doi:10.1111/j.1748-720X.1997.tb01885.x.

This entry was readapted from *Comandé et al. Elgar Encyclopedia of Law and Data Science. Edward Elgar Publishing (2022) ISBN: 978 1 83910 458 9* by Francesca Pratesi, Roberto Pellungrini, and Anna Monreale.

Sustainability

In Brief

Sustainability is an ethical aspect studied in the TAILOR project. Sustainability in AI refers to the application of AI technology (from the power consumption of the hardware to the processing and storage of a very large dataset) while considering problems related to sustainable development. The sustainability in AI can be divided into two branches that should be addressed: AI for Sustainability and Sustainability of AI [1]. The first branch is the most developed branch whose aim is to achieve the United Nations Sustainable Development Goals (SDGs) and where not-for-profit organisation such as AI4Good or AI for Climate are active to address the scope. In this branch we consider AI applications whose objective is to use the algorithm for environmental purposes such as reducing the energy cost of large data centres or better predicting the weather forecast to maximise renewable energy production. The second branch regards the sustainable development of AI/ML itself, concerning how to measure the sustainability of developing and utilising AI models such as the computational cost of training AI algorithms or the amount of CO2 emission.

Abstract

In this part we will cover the main elements that define sustainability in AI systems. Some of them are related to the aspects of AI for sustainability and can be common to computer systems in general, others are related to the sustainability of AI. We will focus on sustainability in AI with more attention to new AI-specific issues and challenges because they are less covered in the literature. We include a taxonomic organisation of terms in the area of AI sustainability and their definition.

Motivation and Background

Given the increasing capabilities and widespread use of artificial intelligence, there is a growing concern about its impact on the environment related to the carbon footprints and the power consumption needed for training, store and developing AI models and algorithms. There is a wide literature regarding the dangers of climate change and the need of modifying the habits of use of the technology by consumers and industries. Plans such as the European Green Deal promulgated by the European Commission has the aim to tackle climate change. AI has the potential to accelerate the efforts of protecting the planet with many applications such as the use of machine learning to optimise the energy consumption efficiency, reducing the CO2 emission, monitoring quality of the air, the water, the biodiversity changes, the vegetation, the forest cover, and preventing natural disasters.

Guidelines

The guidelines include common elements that should be considered in the design and build of a piece of technology while using ML and other emerging technologies. In 2015 the United Nations General Assembly set up the Sustainable Development Goals, a list of 17 interlinked global goals that are intended to be achieved by the year 2030. Not-for-profit organisations such as AI4Good also described a list of guidelines like “Guidelines on the Implementation of Eco-friendly Criteria for AI and other Emerging Technologies” and “Guidelines on the environmental efficiency of machine learning processes in supply chain management” (both guidelines can be found here) with the scope of addressing very important aspects on the use of AI and its development.

Taxonomic Organisation of Terms

The Sustainable Development Goals (SDGs) are a universal call to action to end poverty, protect the planet, and ensure that by 2030 all people enjoy peace and prosperity. They were adopted by the United Nations in 2015 as. They are divided into 17 goals -they recognize that action in one area will affect outcomes in others and that development must balance social, economic and environmental sustainability (see here for further details).

AI is involved in many of these 17 Global Goals, including Sustainable cities and communities and Responsible consumption and production with areas like production optimization, smart cities and green computing.

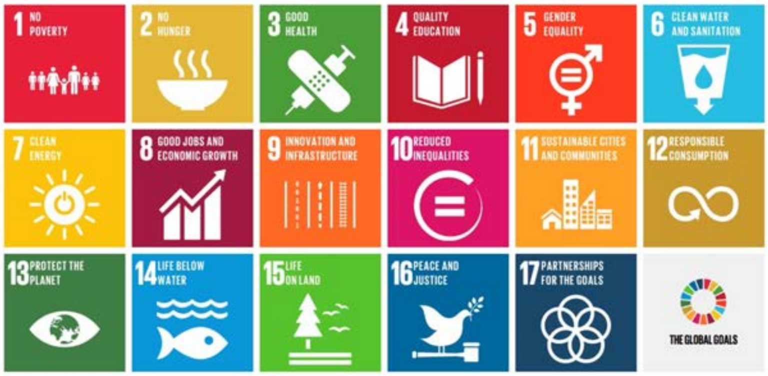


Fig. 19 A possible taxonomy about Sustainability Goals [2].

Main Keywords

- Green AI: The goal of Green AI (also known as *Green IT*, *Green Computing*, or *ICT Sustainability*) is to minimise the negative aspects of IT operations on the environment. To do so, computers and IT products can be designed, manufactured and disposed of in an environmentally-friendly manner.
- Power-aware Computing: Power-aware computing (also called *Energy-aware Computing* and *Energy-efficient Computing*) is part of the Green IT. Power-aware design strategies strive to maximise performance in high-performance systems within power dissipation and power consumption constraints. Reduced power utilisation on a node is one way to reduce the amount of energy required to compute. Lowering the frequency at which the CPU works on one approach to do this. Reduced clock speed, on the other hand, increases the time to solution, posing a potential compromise.
- Cloud Computing: Cloud computing (or *Mesh computing*) is the provision of computing resources (storage and processing power) on demand, without direct user intervention. A large cloud often includes multiple data centres, each housing a different set of functions. The cloud computing model aims to achieve core economies of scale through sharing of resources, taking advantage of a “pay-as-you-go” model that can decrease capital expenditures, but can also result in unforeseen operating expenses for users who are unaware of the concept.
- Edge Computing: Edge Computing (or *Fog Computing*) is a distributed computing paradigm in which processing and data storage are brought closer to the data sources. This should increase response times while also conserving bandwidth. Rather than referring to a single technology, the phrase refers to an architecture.
- Data Centre: A data centre is a structure, a specialised area inside a structure, or a collection of structures used to house computer systems and related components such as telecommunications and storage systems. Because IT operations are so important for business continuity, they usually incorporate redundant or backup components and infrastructure for power, data transmission connections, environmental control (such as air conditioning and fire suppression), and other security systems. A huge data centre is a large-scale activity that consumes the same amount of power as a small town.
- Cradle-to-cradle Design: Cradle-to-cradle Design (also known as *2CC2*, *C2C*, *cradle 2 cradle*, or *regenerative design*) is a biomimetic approach to product and system design that mimics natural processes, in which materials are considered as nutrients flowing in healthy, safe metabolisms.
- Resource Prediction: Resource prediction (also called *Workload Prediction* or *Workload Forecast*) is the estimation of the resources a customer will require in the future to complete his tasks. This concept has a wide variety of application and it is particularly studied in the context of data centres management. When these forecasts are generated, historical and current data are utilised to predict how many resource units, which tools and operative systems and the number of requests are required to accomplish a task.
- Resource Allocation: Cloud computing provides a computing environment where businesses, clients, and projects can lease resources on demand. Both cloud users and providers want to allocate cloud resources efficiently and profitably. These resources are typically scarce, therefore cloud providers must make the best use of them while staying within the confines of the cloud environment and meeting the demands of cloud apps so that they may perform their jobs. The distribution of resources is one of the most important aspects of cloud computing. Its efficiency has a direct impact on the overall performance of the cloud

environment. Cost efficiency, reaction time, reallocation, computing performance, and job scheduling are all key difficulties in resource allocation. Cloud computing users want to do task for the least amount of money feasible.

Bibliography

[[1]] Aimee van Wynsberghe. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3):213–218, 2021.

[[2]] Department of Economic United Nations and Social Affair. Sustainable development. URL: https://sdgs.un.org/goals (visited on 2022-05-02).

Further Recommended Reading

- van Wynsberghe, Aimee. “Sustainable AI: AI for sustainability and the sustainability of AI.” AI and Ethics 1.3 (2021): 213-218.
- Vinuesa, Ricardo, et al. “The role of artificial intelligence in achieving the Sustainable Development Goals.” Nature communications 11.1 (2020): 1-10.
- Graybill, Robert, and Rami Melhem, eds. Power aware computing. Springer Science & Business Media, 2013.
- Schoormann, Thorsten, et al. “Achieving Sustainability with Artificial Intelligence—A Survey of Information Systems Research.” (2021).
- Khakurel, Jayden, et al. “The rise of artificial intelligence under the lens of sustainability.” Technologies 6.4 (2018): 100.
- Nishant, Rohit, Mike Kennedy, and Jacqueline Corbett. “Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda.” International Journal of Information Management 53 (2020): 102104.
- Wu, Carole-Jean, et al. “Sustainable ai: Environmental implications, challenges and opportunities.” arXiv preprint arXiv:2111.00364 (2021).
- Fisher, Douglas H. “Computing and AI for a Sustainable Future.” IEEE intelligent systems 26.6 (2011): 14-18.
- Pedemonte, C. “AI for Sustainability: an overview of AI and the SDGs to contribute to European policy-making.” (2021).
- Galaz, Victor, et al. “Artificial intelligence, systemic risks, and sustainability.” Technology in Society 67 (2021): 101741.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O’Sullivan.

Green AI

Synonyms: Green IT, Green Computing, ICT sustainability.

In Brief

The goal of **Green AI** is to minimise the negative aspects of IT operations on the environment. To do so, computers and IT products can be designed, manufactured and disposed of in an environmentally-friendly manner.

More in Detail

The concept of Green computing practises came into prominence in 1992, when the Environmental Protection Agency (EPA) launched the Energy Star program. The goals of green computing includes the maximisation of energy efficiency, the reduction of energy consumption. There are four aspects that are addressed by green IT:

- *Green use*: use of computers in an eco-friend;y and energy-efficient way by minimising their electricity usage and using their peripheral devices.
- *Green disposal*: disposing of obsolete electronic equipment in an environmentally safe manner by repurposing it or recycling it.
- *Green design*: computers and other devices designed with reduced energy consumption.
- *Green manufacturing*: making computers and other systems with as little waste as possible in order to reduce their environmental impact.

Government regulatory agencies are also actively working to promote green computing principles by enacting a number of voluntary initiatives and rules. The following strategies can be used by average computer users to make their computing more environmentally friendly:

- When you are gone from your computer for a lengthy amount of time, use hibernation or sleep mode.
- Instead of desktop PCs, get energy-efficient laptop computers.
- Make sufficient measures for the safe disposal of electronic trash.
- At the end of the day, turn off computers.
- Rather than buying new printer cartridges, refill them.
- Instead of buying a new computer, consider renovating one that already exists.
- Activate the power management options to keep an eye on your energy usage.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O'Sullivan.

Power-aware Computing

Synonyms: Energy-aware computing, Energy-efficient computing.

In Brief

Power-aware computing is part of the Green AI. Power-aware design strategies strive to maximise performance in high-performance systems within power dissipation and power consumption constraints. Reduced power utilisation on a node is one way to reduce the amount of energy required to compute. Lowering the frequency at which the CPU works on one approach to do this. Reduced clock speed, on the other hand, increases the time to solution, posing a potential compromise.

More in Detail

Low-power design strategies attempt to decrease power or energy consumption in portable equipment in order to fulfil a particular performance or throughput objective. All of the energy used by a system ultimately dissipates and is converted to heat. **Power dissipation** and related thermal issues have an impact on performance, packaging, reliability, environmental impact, and heat removal costs; **power and energy consumption** have an impact on power delivery costs, performance, and reliability, and they are directly related to portable device size and battery life.

Power-aware computing's major purpose is to save energy while routing communications from source to destination. The contemporary period is characterised by wireless networks, in which nodes connect with one another over many hops. In this technology, many data transfer protocols are employed. Different routing techniques will draw more power from the battery. In power-aware computing, power-aware routing metrics have played an essential role.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O'Sullivan.

Cloud Computing

Synonyms: Mesh Computing.

In Brief

Cloud computing is the provision of computing resources (storage and processing power) on demand, without direct user intervention. A large cloud often includes multiple data centres, each housing a different set of functions. The cloud computing model aims to achieve core economies of scale through sharing of resources, taking advantage of a “pay-as-you-go” model that can decrease capital expenditures, but can also result in unforeseen operating expenses for users who are unaware of the concept.

More in Detail

With cloud computing, users are able to utilise any or all of these technologies, without having to know a great deal about them or being an expert in them. As a result, cloud computing reduces overhead and lets users focus on expanding their business rather than on IT problems.

Virtualization is the primary enabling technology for cloud computing. By separating a physical computing device into multiple virtual ones, each of which can easily be managed and used for computing operations, virtualization software allows companies to reduce costs and increase efficiency. By using operating system-level virtualization, idle computing resources can be allocated more efficiently and turned into a scalable system of independent computing devices. Increasing utilisation of the infrastructure through virtualization reduces costs and speeds up IT operations.

In autonomous computing, users can provision resources on-demand by automating a number of steps. Through automation, processes become more efficient, labour costs are reduced, and the possibility of human error is reduced. In the context of Green AI, cloud computing addresses two major challenges - energy consumption and resource consumption.

Cloud computing can significantly lower carbon emissions and energy use thanks to **virtualization**, **dynamic provisioning environment**, **multi-tenancy**, and **green data centre technologies**. Certain on-premises applications can be moved into the cloud by large enterprises and small businesses in order to reduce their energy consumption and carbon emissions. People can buy products and services online without having to travel to physical stores, reducing greenhouse gas emissions associated with travel. One example would be online shopping, where they purchase products online without having to drive and waste fuel to reach the physical stores.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O’Sullivan.

Edge Computing

Synonyms: Fog Computing.

In Brief

Edge computing is a distributed computing paradigm in which processing and data storage are brought closer to the data sources. This should increase response times while also conserving bandwidth. Rather than referring to a single technology, the phrase refers to an architecture.

More in Detail

Edge computing is defined as any form of computer software that provides reduced latency closer to the requests. It can be defined as any computing outside the cloud that happens at the network’s edge, and more particularly in applications where real-time data processing is necessary. While cloud computing works with huge data, edge computing works with immediate data or data created in real-time by sensors or users. Virtualization technology may be used in edge computing to make it easier to deploy and execute a wide range of applications on edge servers.

Edge computing has its roots in content distributed networks, which were developed in the late 1990s to provide web and video content from edge servers located near consumers. In The early 2000s, these networks expanded to host applications and application components at edge servers, leading to the first commercial edge computing services, which hosted applications including dealer locators, shopping carts, real-time data aggregators, and ad insertion engines. Edge and fog computing are examples of new technologies that can help reduce energy use. These technologies enable redistribution computing closer to the user, lowering network energy costs. Furthermore, having fewer data centres reduces the amount of energy consumed in operations such as refrigeration and maintenance.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O’Sullivan.

Data Centre

In Brief

A **data centre** is a structure, a specialised area inside a structure, or a collection of structures used to house computer systems and related components such as telecommunications and storage systems. Because IT operations are so important for business continuity, they usually incorporate redundant or backup components and infrastructure for power, data transmission connections, environmental control (such as air conditioning and fire suppression), and other security systems. A huge data centre is a large-scale activity that consumes the same amount of power as a small town.

More in Detail

Data centres in the field of enterprise IT are meant to serve business applications and operations such as email communication and file sharing, applications for productivity, management of customer relationships, databases and enterprise resource planning, machine learning, AI and big data, communications and collaboration services, as well as virtual desktops. Routers, switches, firewalls, storage systems, servers, and application delivery controllers are all part of the data centre design.

Data centre security is crucial in data centre architecture because these components hold and handle business-critical data and applications. They provide services such as computing resources and network and storage infrastructure. Data centre facilities are energy guzzlers, accounting for between 1.1 and 1.5 percent of total global energy consumption in 2010 [1].

Data centre facilities use to 100 to 200 times more energy than ordinary office buildings, according to the US Department of Energy [2]. From the IT equipment to the HVAC (heating, ventilation and air conditioning) equipment to the actual location, configuration, and construction of the building, an energy efficient data centre design should handle all the energy usage issues contained in a data centre.

The US department of energy has identified five major arrears where energy efficient data centre architecture best practises should be focused:

- IT systems,
- environmental conditions,
- air quality control,
- refrigeration systems,
- and systems involving electricity.

On-site electricity production and waste heat recycling are two more energy-efficient design options recommended.

Data centre design that is energy efficient should assist to better use a data centre’s space while also increasing performance and efficiency.

Bibliography

[[1]] Edward Curry, Bill Guyon, Charles Sheridan, and Brian Donnellan. Developing a sustainable it capability: lessons from Intel's journey. *MIS Quarterly Executive*, 11(2):61–74, 2012.

[[2]] VanGeet, Otto, W. Lintner, and B. Tschudi. FEMP best practises guide for energy-efficient data centre design. 2011. National Renewable Energy Laboratory.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O’Sullivan.

Cradle-to-cradle Design

Synonyms: 2CC2, C2C, cradle 2 cradle, regenerative design.

In Brief

Cradle-to-cradle Design is a biomimetic approach to product and system design that mimics natural processes, in which materials are considered as nutrients flowing in healthy, safe metabolisms.

More in Detail

Cradle to Cradle is a philosophy that views rubbish as an infinite resource and encourages people to do the right thing from the start. It’s about making community and product development work like a healthy ecological system, where all resources are utilised efficiently and in a cyclical manner. Industry, according to C2C, should safeguard and enhance ecosystems and nature’s biological metabolism. It is a comprehensive economic, industrial, and social framework aimed at creating jobs that are efficient and waste-free. The concept may be applied to many facets of human civilization, including urban landscapes, buildings, economy, adn social systems.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O’Sullivan.

Resource Prediction

Synonyms: Workload Prediction, Workload Forecast.

In Brief

Resource prediction is the estimation of the resources a customer will require in the future to complete his tasks. This concept has a wide variety of application and it is particularly studied in the context of data centres management. When these forecasts are generated, historical and current data are utilised to predict how many resource units, which tools and operative systems and the number of requests are required to accomplish a task.

More in Detail

A resource prediction ensures that the resource pool is proportionate to the workload. To accomplish efficient work scheduling and load balancing in cloud computing, accurate resource requests forecast is required. It is vital for a competitive service to ensure that resources are available to fulfil demand as it arises. Cloud computing companies want to preconfigure computers ahead of time in order to deliver a high **Quality of Service** (QoS), which includes low latency, high availability and high dependability. If demand can be precisely forecast, suppliers may expect greater resource usage and a reduction on pre configured but idle computers, in addition to a high Quality of Service.

Cloud service demand, on the other hand, is difficult to forecast due to factors such as variety, size, burst and uncertainty. Service providers would be able to make a principled trade-off between QoS and resource cost if they had an accurate model of demand variance across time. The topic is widely studied in the literature and includes many algorithms and techniques applied over more than a decade, from statistical models to Machine Learning and Deep Learning models.

The benefits of predicting the future demands include better resource utilisation and a reduction of the overall location with the opportunity of serving more customers, which leads to an increase in profit and an overall reduction of energy and maintenance costs, with a possible improvement in environmental impact.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O’Sullivan.

Resource Allocation

In Brief

Cloud computing provides a computing environment where businesses, clients, and projects can lease resources on demand. Both cloud users and providers want to allocate cloud resources efficiently and profitably. These resources are typically scarce, therefore cloud providers must make the best use of them while staying within the confines of the cloud environment and meeting the demands of cloud apps so that they may perform their jobs. The distribution of resources is one of the most important aspects of cloud computing. Its efficiency has a direct impact on the overall performance of the cloud environment. Cost efficiency, reaction time, reallocation, computing performance, and job scheduling are all key difficulties in resource allocation. Cloud computing users want to do task for the least amount of money feasible.

More in Detail

The provision of services and storage space for certain tasks specified by users is referred to as resource allocation. Different resource allocation mechanisms are used to accomplish this. Resource allocation techniques entail combining cloud provider operations when assigning and utilising scarce cloud resources, as well as addressing the demands of cloud applications so that they may accomplish their goals. The two stakeholders in a cloud computing system, cloud consumers and cloud providers, have distinct objectives. Cloud providers encourage customers to utilise as much of their resources as possible in order to increase earnings, whereas consumers have the opposite purpose in mind. They aim to reduce their cloud computing costs while maintaining their performance requirements. There are a variety of ways for achieving a balance between resource allocation and cost. The measures assist in avoiding:

- *Over-provisioning*: happens when the amount of cloud resources available exceeds the amount of resources needed.
- *Under-provisioning*: occurs when the provided resources are insufficient to meet the demand.
- *Resource fragmentation*: is a problem that occurs when a system’s resources are unavailable. The available resources are unable to distribute themselves to the needed users.
- *Resource contention*: when two or more applications in the cloud system want to utilise the same computational resource in the same instance. Cloud computing technology continues to be used by businesses for a variety of purposes, including enhancing productivity, lowering cloud costs, ensuring data security and storing unlimited data. Knowledge of resource allocation is becoming increasingly important for cloud customers that seek to reduce their cloud use expenses. While there are several techniques for distributing resource in the cloud, the most successful method ison that optimises cloud service provider revenu while also ensuring cloud user pleasure.

This entry was written by Andrea Rossi, Andrea Visentin and Barry O’Sullivan.

Index

Here, you can find the list of entries in alphabetical order.

- Attacks on Anonymization Schema
- Attacks on Pseudonymised Data
- Auditing
- Bias
- Black-box Explanation
- Brittleness
- C2C
- Cloud Computing
- Continuous Performance Monitoring
- Cradle 2 cradle
- Cradle-to-cradle Design
- Data Anonymization
- Data Center
- Data Poisoning
- Dependability
- Differential Privacy Models
- Differential Privacy - (\(\epsilon\),\(\delta\))-Differential Privacy
- Differential Privacy - \(\epsilon\)-Differential Privacy
- Differential Privacy - \(\epsilon\)-Indistinguishability
- Distributional Shift
- Dimensions of Explanations
- Grounds of Discrimination
- Distributional Shift
- Direct Behaviour
- Edge Computing
- Energy-aware Computing
- Energy-efficient Computing
- Evaluation
- Equity
- Explanation by Design
- Fair Machine Learning
- Fairness
- Feature Importance
- The Frame Problem
- Fog Computing
- Model Agnostic
- Global Explanations
- Green AI
- Green Computing
- Green IT
- ICT sustainability
- Intended Behaviour
- Justice
- K-anonymity
- Linking Attack
- Local Explanations
- Meaningful Human Control
- Measurement
- Misdirect Behaviour
- Model Agnostic
- Model Specific
- Model Specific
- Negative Side Effects
- Achiving Differential Privacy
- Post-hoc Explanation
- Privacy models
- Provenance Tracking

- Pseudonymization
- Re-identification Attack
- Regenerative Design
- Reliability
- Repeatability
- Replicability
- Reproducibility
- Resource Allocation
- Resource Prediction
- Resource Scheduling
- Robustness
- Saliency Maps
- Security
- Segregation
- Single Tree Approxiamation
- Testing
- Traceability
- Transparency
- Unintended Behaviour
- Wicked Problems
- Workload Forecast
- Workload Prediction

2CC2

Synonyms: Crade-to-cradle design, C2C, cradle 2 cradle, regenerative design.

2CC2 is a biomimetic approach to product and system design that mimics natural processes, in which materials are considered as nutrients flowing in healthy, safe metabolisms.

You can find futher information about 2CC2 here

Accountability

Accountability is an ethical aspect studied in the TAILOR project to ensure that a given actor or actors can render an account of the actions of an AI system. The accountability concept is strictly related to the concept of responsibility.

You can find futher information about Accountability term here

Alignment

Synonyms: (Mis)directed behaviour, (Un)intended behaviour

The goal of AI **alignment** is to ensure that AI systems are aligned with human intentions and values. This first requires determining the normative question of what values or principles we have and what humans really want, collectively or individually, and second, the technical question of how to imbue AI systems with these values and goals.

You can find futher information about Alignment here

Adversarial Attack

Synonyms: Adversarial Input, Adversarial Example.

An **adversarial attack** is any perturbation of the input features or observations of a system (sometimes imperceptible to both humans and the own system) that makes the system fail or take the system to a dangerous state. A prototypical case of an adversarial situation happens with machine learning models, when an external agent maliciously modify input data –often in imperceptible ways– to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a 'stop' sign as a 'speed limit' sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection. A different but related type of adversarial attack is called Data Poisoning, but this involves a malicious compromise of data sources (used for training or testing) at the point of collection and pre-processing.

You can find futher information about Adversarial Attack here

Adversarial Example

Synonyms: Adversarial Input, Adversarial Attack.

An **adversarial attack** is any perturbation of the input features or observations of a system (sometimes imperceptible to both humans and the own system) that makes the system fail or take the system to a dangerous state. A prototypical case of an adversarial situation happens with machine learning models, when an external agent maliciously modify input data –often in imperceptible ways– to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a 'stop' sign as a 'speed limit' sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection. A different but related type of adversarial attack is called Data Poisoning, but this involves a malicious compromise of data sources (used for training or testing) at the point of collection and pre-processing.

You can find futher information about Adversarial Example here

Adversarial Input

Synonyms: Adversarial Attack, Adversarial Example.

An **adversarial input** is any perturbation of the input features or observations of a system (sometimes imperceptible to both humans and the own system) that makes the system fail or take the system to a dangerous state. A prototypical case of an adversarial situation happens with machine learning models, when an external agent maliciously modify input data –often in imperceptible ways– to induce them into misclassification or incorrect prediction. For instance, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon or a 'stop' sign as a 'speed limit' sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection. A different but related type of adversarial attack is called Data Poisoning, but this involves a malicious compromise of data sources (used for training or testing) at the point of collection and pre-processing.

You can find futher information about Adversarial Input here

Ante-hoc Explanation

Synonyms: Explanation by Design, Transparent model.

Ante-hoc explanation means to rely, by design, on a transparent model, instead of providing explanations of an AI model.

You can find futher information about Ante-hoc Explanation here

Assessment

Synonyms: Evaluation, Testing, Measurement.

AI assessment is any activity that estimates attributes as measures— of an AI system or some of its components, abstractly or in particular contexts of operation. These attributes, if well estimated, can be used to explain and predict the behaviour of the system. This can stem from an engineering perspective, trying to understand whether a particular AI system meets the specifications or the intention of their designers, known respectively as **verification** and **validation**. Under this perspective, AI measurement is close to computer systems **testing** (hardware and/or software) and other evaluation procedures in engineering. However, in AI there is an extremely complex adaptive behaviour, and in many cases, with a lack of a written and operational specification. What the systems has to do depends on some constraints and utility functions that have to be optimised, is specified by example (from which the system has to learn a model) or ultimately depends on feedback from the user or the environment (e.g., in the form of rewards).

You can find futher information about Assessment here

Attacks on Anonymization Schema

There are a variety of attacks that involve data privacy. Some of them are very context-specific (for example, there exists attacks on partition-based algorithms, such as deFinetti Attack or Minimality Attack), while other are more general. For example, we can have the *Re-identification Attack*, where the aim is to link a identity in the data with a real identity; *Membership Attack*, where the goal is to understand whether or not a particular individual is present in the considered dataset; *Reconstruction Attack*, where one wants to reconstruct (even partially) a private dataset from public aggregate information.

You can find futher information about \(\epsilon\)-Differential Privacy here

Attacks on Pseudonymised Data

Synonyms: Re-identification Attack, Linking Attack.

A **Attack on Pseudonymised Data** aims to link a certain set of data related to an individual in a dataset (which does not contain direct identifiers) to a real identity, relying on additional information.

You can find futher information about Attacks on Pseudonymised Data here

Auditing

Auditing AI aims to identify and address possible risks and impacts while ensuring robust and trustworthy accountability.

You can find futher information about Auditing here

Bias

Bias refers to an inclination towards or against a particular individual, group, or sub-groups. AI models may inherit biases from training data or introduce new forms of bias.

You can find futher information about Bias here

Black-box Explanation

Synonyms: Post-hoc Explanations.

With a **black-box explanation** we pair the black box model with an interpretation the black box decisions or model, instead of relying, by design, on a transparent model.

You can find futher information about Black-box Explanation here

Brittleness

Synonyms: Robustness.

Brittleness is the degree in which an AI system functions reliably and accurately under harsh conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is therefore the strength of a system's integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, and undesirable reinforcement learning behaviour.

You can find futher information about Brittleness here

C2C

Synonyms: 2CC2, Cradle-to-cradle Design, cradle 2 cradle, regenerative design.

C2C is a biomimetic approach to product and system design that mimics natural processes, in which materials are considered as nutrients flowing in healthy, safe metabolisms.

You can find futher information about C2C here

Cloud Computing

Synonyms: Mesh Computing

Cloud Computing is the provision of computing resources (storage and processing power) on demand, without direct user intervention. A large cloud often includes multiple data centres, each housing a different set of functions. The cloud computing model aims to achieve core economies of scale through sharing of resources, taking advantage of a "pay-as-you-go" model that can decrease capital expenditures, but can also result in unforeseen operating expenses for users who are unaware of the concept.

You can find futher information about Cloud Computing here

Continuous Performance Monitoring

Continuous performance monitoring is the activity to track, log and monitor over time the behaviour and the performance of Artificial Intelligence and Machine Learning models. This activity is particularly relevant after in-production deployment in order to detect any performance drifts and outages of the model.

You can find futher information about Continuous Performance Monitoring term here

Cradle 2 cradle

Synonyms: 2CC2, C2C, Cradle-to-cradle Design, regenerative design.

Cradle 2 cradle is a biomimetic approach to product and system design that mimics natural processes, in which materials are considered as nutrients flowing in healthy, safe metabolisms.

You can find futher information about cradle 2 cradle here

Cradle-to-cradle Design

Synonyms: 2CC2, C2C, cradle 2 cradle, regenerative design.

Cradle-to-cradle Design (also known as *2CC2*, *C2C*, *cradle 2 cradle*, or *regenerative design*) is a biomimetic approach to product and system design that mimics natural processes, in which materials are considered as nutrients flowing in healthy, safe metabolisms.

You can find futher information about Cradle-to-cradle Design here

Data Anonymization

A data subject is considered anonymous if it is reasonably hard to attribute his personal data to him/her.

You can find futher information about Data Anonymization here

Data Center

A **data centre** is a structure, a specialised area inside a structure, or a collection of structures used to house computer systems and related components such as telecommunications and storage systems. Because IT operations are so important for business continuity, they usually incorporate redundant or backup components and infrastructure for power, data transmission connections, environmental control (such as air conditioning and fire suppression), and other security systems. A huge data centre is a large-scale activity that consumes the same amount of power as a small town.

You can find futher information about Data Centers here

Data Poisoning

Data poisoning occurs when an adversary modifies or manipulates part of the dataset upon which a model will be trained, validated, or tested. By altering a selected subset of training inputs, a poisoning attack can induce a trained AI system into curated misclassification, systemic malfunction, and poor performance. An especially concerning dimension of targeted data poisoning is that an adversary may introduce a 'backdoor' into the infected model whereby the trained system functions normally until it processes maliciously selected inputs that trigger error or failure. Data poisoning is possible because data collection and procurement often involves potentially unreliable or questionable sources. When data originates in uncontrollable environments like the internet, social media, or the Internet of Things, many opportunities present themselves to ill-intentioned attackers, who aim to manipulate training examples. Likewise, in third-party data curation processes (such as 'crowdsourced' labelling, annotation, and content identification), attackers may simply handcraft malicious inputs.

You can find futher information about Data Poisoning here

Dependability

Synonyms: Reliability.

The objective of **dependability** is that an AI system behaves exactly as its designers intended and anticipated, over time. A reliable system adheres to the specifications it was programmed to carry out at any time. Reliability is therefore a measure of consistency of operation and can establish confidence in the safety of a system based upon the dependability with which it operationally conforms to its intended functionality.

You can find futher information about Dependability here

Differential Privacy Models

Differential privacy implies that adding or deleting a single record does not significantly affect the result of any analysis.

You can find futher information about Differential Privacy Models here

(\(\epsilon\),\(\delta\))-Differential Privacy

A relaxed version of *Differential Privacy*, named **(\(\epsilon\),\(\delta\))-Differential Privacy**, allows a little privacy loss (\(\delta\)) due to a variation in the output distribution for the privacy mechanism.

You can find futher information about (\(\epsilon\),\(\delta\))-Differential Privacy here

\(\epsilon\)-Differential Privacy

Synonyms: \(\epsilon\)-indistinguishability.

\(\epsilon\)-Differential Privacy is the simpler form of *Differential Privacy*, where \(\epsilon\) represents the level of privacy guarantee.

You can find futher information about \(\epsilon\)-Differential Privacy here

\(\epsilon\)-Indistinguishability

Synonyms: \(\epsilon\) Differential Privacy.

\(\epsilon\)-Indistinguishability is the simpler form of *Differential Privacy*, where \(\epsilon\) represents the level of privacy guarantee.

You can find futher information about \(\epsilon\)-Indistinguishability here

Distributional Shift

Synonyms: Distributional Shift.

Once trained, most machine learning systems operate on static models of the world that have been built from historical data which have become fixed in the systems' parameters. This freezing of the model before it is released 'into the wild' makes its accuracy and reliability especially vulnerable to changes in the underlying distribution of data. When the historical data that have crystallised into the trained model's architecture cease to reflect the population concerned, the model's mapping function will no longer be able to accurately and reliably transform its inputs into its target output values. These systems can quickly become prone to error in unexpected and harmful ways. In all cases, the system and the operators must remain vigilant to the potentially rapid concept drifts that may occur in the complex, dynamic, and evolving environments in which your AI project will intervene. Remaining aware of these transformations in the data is crucial for safe AI.

You can find futher information about Data Shift here

Dimensions of Explanations

Dimensions of explanations are useful to analyze the interpretability of AI systems and to classify the explanation method.

You can find futher information about Dimensions of Explanations here

Grounds of Discrimination

International and national laws prohibit **discriminating on some explicitly defined grounds**, such as race, sex, religion, etc. They can be considered in isolation, or interacting, giving rise to multiple discrimination and intersectional discrimination.

You can find futher information about Discrimination here

Distributional Shift

Synonyms: Data Shift.

Once trained, most machine learning systems operate on static models of the world that have been built from historical data which have become fixed in the systems' parameters. This freezing of the model before it is released 'into the wild' makes its accuracy and reliability especially vulnerable to changes in the underlying distribution of data. When the historical data that have crystallised into the trained model's architecture cease to reflect the population concerned, the model's mapping function will no longer be able to accurately and reliably transform its inputs into its target output values. These systems can quickly become prone to error in unexpected and harmful ways. In all cases, the system and the operators must remain vigilant to the potentially rapid concept drifts that may occur in the complex, dynamic, and evolving environments in which your AI project will intervene. Remaining aware of these transformations in the data is crucial for safe AI.

You can find futher information about Distributional Shift here

Direct Behaviour

Synonyms: Aligment, (Un)intended behaviour

The goal of AI **direct behaviour** is to ensure that AI systems are aligned with human intentions and values. This first requires determining the normative question of what values or principles we have and what humans really want, collectively or individually, and second, the technical question of how to imbue AI systems with these values and goals.

You can find futher information about Direct Behaviour here

Edge Computing

Synonyms: Fog Computing.

Edge Computing is a distributed computing paradigm in which processing and data storage are brought closer to the data sources. This should increase response times while also conserving bandwidth. Rather than referring to a single technology, the phrase refers to an architecture.

You can find futher information about Edge Computing here

Energy-aware Computing

Synonyms: Power-aware computing, Energy-efficient computing.

Energy-aware computing is part of the Green IT. Power-aware design strategies strive to maximise performance in high-performance systems within power dissipation and power consumption constraints. Reduced power utilisation on a node is one way to reduce the amount of energy required to compute. Lowering the frequency at which the CPU works on one approach to do this. Reduced clock speed, on the other hand, increases the time to solution, posing a potential compromise.

You can find futher information about Energy-aware Computing here

Energy-efficient Computing

Synonyms: Power-aware computing, Energy-aware computing.

Energy-efficient computing is part of the Green IT. Power-aware design strategies strive to maximise performance in high-performance systems within power dissipation and power consumption constraints. Reduced power utilisation on a node is one way to reduce the amount of energy required to compute. Lowering the frequency at which the CPU works on one approach to do this. Reduced clock speed, on the other hand, increases the time to solution, posing a potential compromise.

You can find futher information about Energy-efficient Computing here

Evaluation

Synonyms: Assessment, Testing, Measurement.

AI evaluation is any activity that estimates attributes as measures— of an AI system or some of its components, abstractly or in particular contexts of operation. These attributes, if well estimated, can be used to explain and predict the behaviour of the system. This can stem from an engineering perspective, trying to understand whether a particular AI system meets the specifications or the intention of their designers, known respectively as **verification** and **validation**. Under this perspective, AI measurement is close to computer systems **testing** (hardware and/or software) and other evaluation procedures in engineering. However, in AI there is an extremely complex adaptive behaviour, and in many cases, with a lack of a written and operational specification. What the systems has to do depends on some constraints and utility functions that have to be optimised, is specified by example (from which the system has to learn a model) or ultimately depends on feedback from the user or the environment (e.g., in the form of rewards).

You can find futher information about Evaluation here

Equity

Forms of bias that count as discrimination against social groups or individuals should be avoided, both from legal and ethical perspectives. Discrimination can be direct or indirect, intentional or unintentional.

You can find futher information about Equity here

Explanation by Design

Synonyms: Ante-hoc Explanation, Transparent model.

Explanation by Design means to rely, by design, on a transparent model, instead of providing explanations of an AI model.

You can find futher information about Explanation by Design here

Fair Machine Learning

Fair Machine Learning models take into account the issues of bias and fairness. Approaches can be categorized as pre-processig, which transform the input data, as in-processing, which modify the learning algorithm, and post-processing, which alter models' internals or their decisions.

You can find futher information about Fair Machine Learning here

Fairness

The term **fairness** is defined as the quality or state of being fair; or a lack of favoritism towards one side. The notions of fairness, and quantitative measures of them (fairness metrics), can be distinguished based on the focus on individuals, groups and sub-groups.

You can find futher information about Fairness here

Feature Importance

The **feature importance** technique provides a score, representing the "importance", for all the input features for a given AI model, i.e., a higher importance means that the corresponding feature will have a larger effect on the model.

You can find futher information about Feature Importance here

The Frame Problem

The **frame problem** is the challenge of knowing and modeling the relevant features and context of situations, and getting an agent to act on those without consideration all the irrelevant facts as well.

You can find futher information about The Frame Problem term here

Fog Computing

Synonyms: Edge Computing.

Fog Computing is a distributed computing paradigm in which processing and data storage are brought closer to the data sources. This should increase response times while also conserving bandwidth. Rather than referring to a single technology, the phrase refers to an architecture.

You can find futher information about Fog Computing here

Model Agnostic

Synonyms: Model Agnostic Explanation.

We distinguish between **model-specific** or **model-agnostic** explanation method depending on whether the technique adopted to retrieve the explanation acts on a particular model adopted by an AI system, or can be used on any type of AI.

You can find futher information about Model Agnostic term here

Global Explanations

Global explanation is an explanation that allows understanding the whole logic of a model used by an AI system.

You can find futher information about Global Explanations here

Green AI

Synonyms: Green IT, Green Computing, ICT sustainability.

The goal of **Green AI** is to minimise the negative aspects of IT operations on the environment. To do so, computers and IT products can be designed, manufactured and disposed of in an environmentally-friendly manner.

You can find futher information about Green AI here

Green Computing

Synonyms: Green IT, Green AI, ICT sustainability.

The goal of **Green Computing** is to minimise the negative aspects of IT operations on the environment. To do so, computers and IT products can be designed, manufactured and disposed of in an environmentally-friendly manner.

You can find futher information about Green Computing here

Green IT

Synonyms: Green AI, Green Computing, ICT sustainability.

The goal of **Green IT** is to minimise the negative aspects of IT operations on the environment. To do so, computers and IT products can be designed, manufactured and disposed of in an environmentally-friendly manner.

You can find futher information about Green IT here

ICT sustainability

Synonyms: Green IT, Green Computing, Green AI.

The goal of **ICT sustainability** is to minimise the negative aspects of IT operations on the environment. To do so, computers and IT products can be designed, manufactured and disposed of in an environmentally-friendly manner.

You can find futher information about ICT sustainability here

Intended Behaviour

Synonyms: (Mis)directed behaviour, Alignement

The goal of AI **intended behaviour** is to ensure that AI systems are aligned with human intentions and values. This first requires determining the normative question of what values or principles we have and what humans really want, collectively or individually, and second, the technical question of how to imbue AI systems with these values and goals.

You can find futher information about Intended Behaviour here

Justice

Justice encompasses three different perspectives: (1) *fairness* understood as the fair treatment of people, (2) *rightness* as the quality of being fair or reasonable, and (3) a legal system, the scheme or system of law. Justice can be distinguished between *substantive* and *procedural*.

You can find futher information about Justice here

K-anonymity

k-anonimity (and the whole family of anonymity by **indistinguishability models**) is based on comparison among individuals present in data, and it aims to make each individual so similar as to be indistinguishable from at least *k-1* others.

You can find futher information about K-anonymity here

Linking Attack

Synonyms: Re-identification Attack, Attack on Pseudonymised Data.

Linking Attack attack aims to link a certain set of data related to an individual in a dataset (which does not contain direct identifiers) to a real identity, relying on additional information.

You can find futher information about Linking Attack here

Local Explanations

Local explanation is an explanation that refers to a specific case, i.e., only a single decision is interpretable.

You can find futher information about Local Explanations here

Meaningful Human Control

Meaningful human control is the notion that aims to generalize the traditional concept of operational control over technological artifacts to artificial intelligent systems. It implies that artificial systems should not make morally consequential decisions on their own, without appropriate control from responsible humans.

You can find futher information about Meaningful Human Control here

Measurement

Synonyms: Assessment, Testing, Evaluation.

AI measurement is any activity that estimates attributes as measures— of an AI system or some of its components, abstractly or in particular contexts of operation. These attributes, if well estimated, can be used to explain and predict the behaviour of the system. This can stem from an engineering perspective, trying to understand whether a particular AI system meets the specifications or the intention of their designers, known respectively as **verification** and **validation**. Under this perspective, AI measurement is close to computer systems **testing** (hardware and/or software) and other evaluation procedures in engineering. However, in AI there is an extremely complex adaptive behaviour, and in many cases, with a lack of a written and operational specification. What the systems has to do depends on some constraints and utility functions that have to be optimised, is specified by example (from which the system has to learn a model) or ultimately depends on feedback from the user or the environment (e.g., in the form of rewards).

You can find futher information about Measurement here

Mesh Computing

Synonyms: Cloud Computing

Mesh Computing is the provision of computing resources (storage and processing power) on demand, without direct user intervention. A large cloud often includes multiple data centres, each housing a different set of functions. The cloud computing model aims to achieve core economies of scale through sharing of resources, taking advantage of a “pay-as-you-go” model that can decrease capital expenditures, but can also result in unforeseen operating expenses for users who are unaware of the concept.

You can find futher information about Mesh Computing here

Misdirect Behaviour

Synonyms: Aligment, (Un)intended behaviour

The goal of AI **direct behaviour** is to ensure that AI systems are aligned with human intentions and values. This first requires determining the normative question of what values or principles we have and what humans really want, collectively or individually, and second, the technical question of how to imbue AI systems with these values and goals.

You can find futher information about Misdirect Behaviour here

Model Agnostic

Synonyms: Generalizable Explanation.

We distinguish between **model-specific** or **model-agnostic** explanation method depending on whether the technique adopted to retrieve the explanation acts on a particular model adopted by an AI system, or can be used on any type of AI.

You can find futher information about Model Agnostic term here

Model Specific

Synonyms: Not Generalizable Explanation.

We distinguish between **model-specific** or **model-agnostic** explanation method depending on whether the technique adopted to retrieve the explanation acts on a particular model adopted by an AI system, or can be used on any type of AI.

You can find futher information about Model Specific term here

Negative Side Effects

Negative side effects are an important safety issue in AI system that considers all possible unintended harm that is caused as a secondary effect of the AI system's operation. An agent can disrupt or break other systems around, or damage third parties, including humans, or can exhaust resources, or a combination of all this. This usually happens because many things the system should not do are not included in its specification. In the case of AI systems, this is even more poignant as written specifications are usually replaced by an optimisation or loss function, in which it is even more difficult to express these things the system should not do, as they frequently rely on 'common sense'.

You can find futher information about Negative Side Effects here

Model Specific

Synonyms: Not Generalizable Explanation.

We distinguish between **model-specific** or **model-agnostic** explanation method depending on whether the technique adopted to retrieve the explanation acts on a particular model adopted by an AI system, or can be used on any type of AI.

You can find futher information about Model Specific term here

Achiving Differential Privacy

Differential privacy guarantees can be provided by **perturbation mechanisms** aim at randomizing the output distributions of functions in order to provide privacy guarantees.

You can find futher information about Achiving Differential Privacy here

Post-hoc Explanation

Synonyms: Black-box Explanations.

With a **post-hoc explanation** we pair the black box model with an interpretation the black box decisions or model, instead of relying, by design, on a transparent model.

You can find futher information about Post-hoc Explanation here

Power-aware Computing

Synonyms: Energy-aware computing, Energy-efficient computing.

Power-aware computing is part of the Green IT. Power-aware design strategies strive to maximise performance in high-performance systems within power dissipation and power consumption constraints. Reduced power utilisation on a node is one way to reduce the amount of energy required to compute. Lowering the frequency at which the CPU works on one approach to do this. Reduced clock speed, on the other hand, increases the time to solution, posing a potential compromise.

You can find futher information about Power-aware Computing here

Privacy models

There are essentially two families of models, based on different goals and mechanisms: *anonymity by randomization* (where the most recent paradigm is *Differential Privacy*) and *anonymity by indistinguishability* (whose most famous example is *k-anonymity*).

You can find futher information about Privacy models here

Provenance Tracking

Provenance tracking represents the tracking of "information that describes the production process of an end product, which can be anything from a piece of data to a physical object. […] Essentially, provenance can be seen as meta-data that, instead of describing data, describes a production process."

You can find futher information about Provenance Tracking term here

Pseudonymization

Pseudonymisation aims to substitute one or more identifiers that link(s) the identity of an individual to its data with a surrogate value, called **pseudonym** or **token**.

You can find futher information about Data Pseudonymization here

Re-identification Attack

Synonyms: Linking Attack, Attack on Pseudonymised Data.

Re-identification attack aims to link a certain set of data related to an individual in a dataset (which does not contain direct identifiers) to a real identity, relying on additional information.

You can find futher information about Re-identification Attack here

Regenerative Design

Synonyms: 2CC2, C2C, cradle 2 cradle, Cradle-to-cradle Design.

Regenerative Design is a biomimetic approach to product and system design that mimics natural processes, in which materials are considered as nutrients flowing in healthy, safe metabolisms.

You can find futher information about Regenerative Design here

Reliability

Synonyms: Dependability.

The objective of **reliability** is that an AI system behaves exactly as its designers intended and anticipated, over time. A reliable system adheres to the specifications it was programmed to carry out at any time. Reliability is therefore a measure of consistency of operation and can establish confidence in the safety of a system based upon the dependability with which it operationally conforms to its intended functionality.

You can find futher information about Reliability here

Repeatability

Synonyms: Reproducibility, Replicability.

Repeatability is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.

You can find futher information about Repeatability here

Replicability

Synonyms: Reproducibility, Repeatability.

Replicability is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.

You can find futher information about Replicability here

Reproducibility

Synonyms: Repeatability, Replicability.

Reproducibility is the ability of independent investigators to draw the same conclusions from an experiment by following the documentation shared by the original investigators.

You can find futher information about Reproducibility here

Resource Allocation

Cloud computing provides a computing environment where businesses, clients, and projects can lease resources on demand. Both cloud users and providers want to allocate cloud resources efficiently and profitably. These resources are typically scarce, therefore cloud providers must make the best use of them while staying within the confines of the cloud environment and meeting the demands of cloud apps so that they may perform their jobs. The distribution of resources is one of the most important aspects of cloud computing. Its efficiency has a direct impact on the overall performance of the cloud environment. Cost efficiency, reaction time, reallocation, computing performance, and job scheduling are all key difficulties in resource allocation. Cloud computing users want to do task for the least amount of money feasible.

You can find futher information about Resource Allocation here

Resource Prediction

Synonyms: Workload Forecast, Workload Prediction.

Resource prediction is the estimation of the resources a customer will require in the future to complete his tasks. This concept has a wide variety of application and it is particularly studied in the context of data centres management. When these forecasts are generated, historical and current data are utilised to predict how many resource units, which tools and operative systems and the number of requests are required to accomplish a task.

You can find futher information about Resource Prediction here

Resource Scheduling

Cloud computing provides a computing environment where businesses, clients, and projects can lease resources on demand. Both cloud users and providers want to allocate cloud resources efficiently and profitably. These resources are typically scarce, therefore cloud providers must make the best use of them while staying within the confines of the cloud environment and meeting the demands of cloud apps so that they may perform their jobs. The distribution of resources is one of the most important aspects of cloud computing. Its efficiency has a direct impact on the overall performance of the cloud environment. Cost efficiency, reaction time, reallocation, computing performance, and job scheduling are all key difficulties in resource allocation. Cloud computing users want to do task for the least amount of money feasible.

You can find futher information about Resource Scheduling here

Robustness

Synonyms: Brittleness.

Robustness is the degree in which an AI system functions reliably and accurately under harsh conditions. These conditions may include adversarial intervention, implementer error, or skewed goal-execution by an automated learner (in reinforcement learning applications). The measure of robustness is therefore the strength of a system's integrity and the soundness of its operation in response to difficult conditions, adversarial attacks, perturbations, data poisoning, and undesirable reinforcement learning behaviour.

You can find futher information about Robustness here

Saliency Maps

Saliency maps are explanations used on image classification tasks. A saliency map is an image where each pixel's color represents a value modeling the importance of that pixel in the original image (i.e., the one given in input to the explainer) for the prediction.

You can find futher information about Saliency Maps here

Security

The goal of **security** encompasses the protection of several operational dimensions of an AI system when confronted with possible attacks, trying to take control of the system or having access to design, operational or personal information. A secure system is capable of maintaining the integrity of the information that constitutes it. This includes protecting its architecture from the unauthorised modification or damage of any of its component parts. A secure system also keeps confidential and private information protected even under hostile or adversarial conditions.

You can find futher information about Security here

Segregation

Social segregation refers to the separation of groups on the grounds of personal or cultural traits. Separation can be physical (e.g., in schools or neighborhoods) or virtual (e.g., in social networks).

You can find futher information about Segregation here

Single Tree Approxiamation

The **single tree appoximation** is an approach that aims at building a decision tree to approximate the behavior of a black box, typically a neural network.

You can find futher information about Single Tree Approxiamation here

Testing

Synonyms: Assessment, Evaluation, Measurement.

AI evaluation is any activity that estimates attributes as measures— of an AI system or some of its components, abstractly or in particular contexts of operation. These attributes, if well estimated, can be used to explain and predict the behaviour of the system. This can stem from an engineering perspective, trying to understand whether a particular AI system meets the specifications or the intention of their designers, known respectively as **verification** and **validation**. Under this perspective, AI measurement is close to computer systems **testing** (hardware and/or software) and other evaluation procedures in engineering. However, in AI there is an extremely complex adaptive behaviour, and in many cases, with a lack of a written and operational specification. What the systems has to do depends on some constraints and utility functions that have to be optimised, is specified by example (from which the system has to learn a model) or ultimately depends on feedback from the user or the environment (e.g., in the form of rewards).

You can find futher information about Testing here

Traceability

Traceability can be defined as the need to maintain a complete and clear documentation of the data, processes, artefacts and actors involved in the entire lifecycle of an AI model, starting from its design and ending with its production serving.

You can find futher information about Traceability here

Transparency

Synonyms: Explanation by Design, Ante-hoc Explanation.

Transparency means to rely, by design, on a transparent model, instead of providing explanations of an AI model.

You can find futher information about Transparency here

Unintended Behaviour

Synonyms: (Mis)directed behaviour, Alignement

The goal of AI **intended behaviour** is to ensure that AI systems are aligned with human intentions and values. This first requires determining the normative question of what values or principles we have and what humans really want, collectively or individually, and second, the technical question of how to imbue AI systems with these values and goals.

You can find futher information about Unintended Behaviour here

Wicked Problems

A class of problems for which science provides insufficient or inappropriate resolution.

You can find futher information about Wicked Problems term here

Workload Forecast

Synonyms: Resource Prediction, Workload Prediction.

Workload Forecast is the estimation of the resources a customer will require in the future to complete his tasks. This concept has a wide variety of application and it is particularly studied in the context of data centres management. When these forecasts are generated, historical and current data are utilised to predict how many resource units, which tools and operative systems and the number of requests are required to accomplish a task.

You can find futher information about Workload Forecast here

Workload Prediction

Synonyms: Resource Prediction, Workload Forecast.

Workload prediction is the estimation of the resources a customer will require in the future to complete his tasks. This concept has a wide variety of application and it is particularly studied in the context of data centres management. When these forecasts are generated, historical and current data are utilised to predict how many resource units, which tools and operative systems and the number of requests are required to accomplish a task.

You can find futher information about Workload Prediction here

Di TAILOR WP3 members; see here for the complete list of contributors.