Contents lists available at ScienceDirect

Explainable Artiﬁcial Intelligence (XAI): Concepts, taxonomies,

opportunities and challenges toward responsible AI

a TECNALIA, Derio 48160, Spain

b ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France

c University of the Basque Country (UPV/EHU), Bilbao 48013, Spain

d Basque Center for Applied Mathematics (BCAM), Bilbao 48009, Bizkaia, Spain

g DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, Granada 18071, Spain

h Telefonica, Madrid 28050, Spain

Explainable Artiﬁcial Intelligence

Deep Learning

Fairness

In the last few years, Artiﬁcial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately,

Machine Learning, the entire community stands in front of the barrier of explainability, an inherent problem of

in the last hype of AI (namely, expert systems and rule based models). Paradigms underlying this problem fall

within the so-called eXplainable AI (XAI) ﬁeld, which is widely acknowledged as a crucial feature for the practical

deployment of AI models. The overview presented in this article examines the existing literature and contributions

already done in the ﬁeld of XAI, including a prospect toward what is yet to be reached. For this purpose we

summarize previous eﬀorts made to deﬁne explainability in Machine Learning, establishing a novel deﬁnition of

explainable Machine Learning that covers such prior conceptual propositions with a major focus on the audience

for which the explainability is sought. Departing from this deﬁnition, we propose and discuss about a taxonomy

of recent contributions related to the explainability of diﬀerent Machine Learning models, including those aimed

at explaining Deep Learning methods for which a second dedicated taxonomy is built and examined in detail.

This critical literature analysis serves as the motivating background for a series of challenges faced by XAI,

such as the interesting crossroads of data fusion and explainability. Our prospects lead toward the concept of

Responsible Artiﬁcial Intelligence , namely, a methodology for the large-scale implementation of AI methods in real

organizations with fairness, model explainability and accountability at its core. Our ultimate goal is to provide

newcomers to the ﬁeld of XAI with a thorough taxonomy that can serve as reference material in order to stimulate

beneﬁts of AI in their activity sectors, without any prior bias for its lack of interpretability.

Artiﬁcial Intelligence (AI) lies at the core of many activity sectors

of AI trace back to several decades ago, there is a clear consensus on

virtue of these capabilities that AI methods are achieving unprecedented

∗ Corresponding author at TECNALIA. P. Tecnologico, Ed. 700. 48170 Derio (Bizkaia), Spain.

E-mail address: javier.delser@tecnalia.com (J. Del Ser). levels of performance when learning to solve increasingly complex com-

human society [2] . The sophistication of AI-powered systems has lately

sions are furnished by AI methods [3] .

Available online 26 December 2019

Fig. 1. Evolution of the number of total publications whose title, abstract and/or keywords refer to the ﬁeld of XAI during the last years. Data retrieved from Scopus®

(December 10th, 2019) by using the search terms indicated in the legend when querying this database. It is interesting to note the latent need for interpretable AI

to explain AI models has permeated throughout the research community.

While the very ﬁrst AI systems were easily interpretable, the last

Neural Networks (DNNs). The empirical success of Deep Learning (DL)

mand for transparency is increasing from the various stakeholders in AI

legitimate, or that simply do not allow obtaining detailed explanations

ing demand for ethical AI [3] . It is customary to think that by focusing

detect, and consequently, correct from bias in the training dataset.

tion of AI systems, eXplainable AI (XAI) [7] proposes creating a suite of

ML techniques that 1) produce more explainable models while main-

taining a high level of learning performance (e.g., prediction accuracy),

manage the emerging generation of artiﬁcially intelligent partners. XAI

Fig. 1 displays the rising trend of contributions on XAI and related

surveys [8,10,13–17] summarize the upsurge of activity in XAI across

sectors and disciplines, this overview aims to cover the creation of a

scrutiny and understanding of the ﬁeld of XAI methods. Furthermore, we

pose intriguing thoughts around the explainability of AI models in data

Artiﬁcial Intelligence, term by which we refer to a series of AI princi-

ples to be necessarily met when deploying AI in real applications. As

we will later show in detail, model explainability is among the most

1. Grounded on a ﬁrst elaboration of concepts and terms used in XAI-

related research, we propose a novel deﬁnition of explainability that

plaining a ML model. We also elaborate on the diverse purposes

sought when using XAI techniques, from trustworthiness to privacy

awareness, which round up the claimed importance of purpose and

targeted audience in model explainability.

post-hoc explainability, namely, the explanation of ML models that

Fig. 2. Diagram showing the diﬀerent purposes of explainability in ML models sought by diﬀerent audience proﬁles. Two goals occur to prevail across them: need

3. We thoroughly analyze the literature on XAI and related concepts

the explainability of ML models using the previously made distinc-

tion between transparency and post-hoc explainability, including

shallow ) learning models. The second taxonomy deals with XAI meth-

ods suited for the explanation of Deep Learning models, using clas-

4. We enumerate a series of challenges of XAI that still remain insuf-

around the concepts and metrics to evaluate the explainability of ML

our prospects toward the implications of XAI techniques in regards

and other areas intersecting with explainability.

poses the systematic adoption of several AI principles for AI models

to be of practical use. In addition to explainability, the guidelines

behind Responsible AI establish that fairness, accountability and pri-

vacy should also be considered when implementing AI models in real

6. Since Responsible AI blends together model explainability and pri-

beneﬁts and risks of XAI techniques in scenarios dealing with sen-

and governance demands more eﬀorts to assess the role of XAI in

of XAI in terms of privacy and security under diﬀerent data fusion

The remainder of this overview is structured as follows: ﬁrst,

and concepts revolving around explainability and interpretability in AI,

ML models from the XAI perspective. Sections 3 and 4 proceed by re-

viewing recent ﬁndings on XAI for ML models (on transparent models

and post-hoc techniques respectively) that comprise the main division in

the aforementioned taxonomy. We also include a review on hybrid ap- proaches among the two, to attain XAI. Beneﬁts and caveats of the syn-

outlook aimed at engaging the community around this vibrant research

2. Explainability: What, why, what for and how?

establish a common point of understanding on what the term explain-

ability stands for in the context of AI and, more speciﬁcally, ML. This

argue why explainability is an important issue in AI and ML (why? what

for?) and to introduce the general classiﬁcation of XAI approaches that

is the interchangeable misuse of interpretability and explainability in

explainability can be viewed as an active characteristic of a model, de-

clarifying or detailing its internal functions.

ethical AI and XAI communities.

how the model works –without any need for explaining its internal

•Interpretability : It is deﬁned as the ability to explain or to provide

•Explainability : Explainability is associated with the notion of expla-

essential concept in XAI. Both transparency and interpretability are

understand the knowledge contained in the model. All in all, under-

understandability. This is the reason why the deﬁnition of XAI given

cept of audience the cornerstone of XAI, as we next elaborate in further

detail.

AI. It appears from the literature that there is not yet a common point of

understanding on what interpretability or explainability are. However,

many contributions claim the achievement of interpretable models and

techniques that empower explainability.

Explainable Artiﬁcial Intelligence (XAI) given by D. Gunning in [7] :

“XAI will create a suite of machine learning techniques that enables hu-

sider other purposes motivating the need for interpretable AI models,

such as causality, transferability, informativeness, fairness and conﬁ-

nition of explainability in AI still slips from our ﬁngers. A broader re-

formulation of this deﬁnition (e.g. “An explainable Artiﬁcial Intelligence

is one that produces explanations about its functioning ”) would fail to fully

an explanation is “the details or reasons that someone gives to make some-

this can be rephrased as: “the details or reasons a model gives to make its

ambiguities can be pointed out. First, the details or the reasons used

to explain, are completely dependent of the audience to which they are

the explainability of the model on the audience. To this end, a reworked

Given a certain audience, explainability refers to the details and reasons

Since explaining, as argumenting, may involve weighting, comparing

arguments [28] , explainability might convey us into the realm of cogni-

ternals of a model can be explained could be tackled objectively. Any

should be considered as an XAI approach. How big this leap is in terms

of complexity or simplicity will correspond to how explainable the re-

sulting model is. An underlying problem that remains unsolved is that

the interpretability gain provided by such XAI approaches may not be

straightforward to quantify: for instance, a model simpliﬁcation can be

tiﬁcation of the improvements gained in terms of interpretability. The

derivation of general metrics to assess the quality of XAI approaches re-

main as an open challenge that should be under the spotlight of the ﬁeld

Explainability is linked to post-hoc explainability since it covers the

techniques used to convert a non-interpretable model into a explain-

able one. In the remaining of this manuscript, explainability will be

considered as the main design objective, since it represents a broader

concept. A model can be explained, but the interpretability of the

Bearing these observations in mind, explainable AI can be deﬁned as

Given an audience, an explainable Artiﬁcial Intelligence is one that pro-

duces details or reasons to make its functioning clear or easy to under-

targeted by XAI techniques for the model at hand reverts on diﬀerent

Goals pursued in the reviewed literature toward reaching explainability, and their main target audience.

XAI Goal Main target audience ( Fig. 2 ) References

Trustworthiness Domain experts, users of the model affected by decisions [5,10,24,32–37]

Causality Domain experts, managers and executive board members,

Transferability Domain experts, data scientists [5,21,26,30,32,37–39,44–85]

Conﬁdence Domain experts, developers, managers, regulatory entities/agencies [5,35,45,46,48,54,61,72,88,89,96,108,117,119,155]

Fairness Users affected by model decisions, regulatory entities/agencies [5,24,35,45,47,99–101,120,121,128,156–158]

Interactivity Domain experts, users affected by model decisions [37,50,59,65,67,74,86,124]

As stated in the introduction, explainability is one of the main barri-

ers AI is facing nowadays in regards to its practical implementation. The

inability to explain or to fully understand the reasons by which state-

The second axis is that of knowledge. AI has helped research across

able data has largely beneﬁted from the adoption of AI and ML tech-

though for certain disciplines this might be the fair case, science and

goals motivating the search for explainable AI models.

The research activity around XAI has so far exposed diﬀerent goals

to draw from the achievement of an explainable model. Almost none

scribe what an explainable model should compel. However, all these

exercise of ML explainability is performed. Unfortunately, scarce con-

these XAI goals, so as to settle a ﬁrst classiﬁcation criteria for the full

thiness as the primary aim of an explainable AI model [31,32] . How-

ever, declaring a model as explainable as per its capabilities of in-

model explainability. Trustworthiness might be considered as the

a given problem. Although it should most certainly be a property

of any explainable model, it does not imply that every trustwor-

thy model can be considered explainable on its own, nor is trust-

worthiness a property easy to quantify. Trust might be far from being the only purpose of an explainable model since the relation

pose for achieving explainability. However, as seen in Table 1 , they

to XAI.

•Causality: Another common goal for explainability is that of ﬁnding

causality among data variables. Several authors argue that explain-

involves correlation, so an explainable ML model could validate the

intuition of possible causal relationships within the available data.

Again, Table 1 reveals that causality is not among the most impor-

•Transferability: Models are always bounded by constraints that

should allow for their seamless transferability. This is the main rea-

son why a training-testing approach is used when dealing with ML

problems [162,163] . Explainability is also an advocate for transfer-

explainable model, but again, not every transferable model should

be considered as explainable. As observed in Table 1 , the amount of

papers stating that the ability of rendering a model explainable is

explainability.

pitfalls. For this purpose, explainable ML models should give infor-

However, other targets for explainability may include a given example, the output classes or the dataset itself.

consider explaining the antecedent. This is the most used argument

reaching explainable models.

is expected. The methods to maintain conﬁdence under control are

is a must-have when drawing interpretations from a certain model.

are not stable. Hence, an explainable model should contain informa-

•Fairness: From a social standpoint, explainability can be considered

as the capacity to reach and guarantee fairness in ML models. In a

certain literature strand, an explainable ML model suggests a clear

visualization of the relations aﬀecting a result, allowing for a fairness

objective of XAI is highlighting bias in the data a model was exposed

in ﬁelds that involve human lives, hence explainability should be

considered as a bridge to avoid the unfair or unethical use of algo-

for explainability as the property that allows end users to get more

involved in the process of improving and developing a certain ML

model [37,86] . It seems clear that explainable models will ease the

an explainable ML model. Once again, this goal is related to ﬁelds

the byproducts enabled by explainability in ML models is its ability

may entail a privacy breach. Contrarily, the ability to explain the

inner relations of a trained model by non-authorized third parties

to its criticality in sectors where XAI is foreseen to play a crucial

scope of the reviewed papers. All these goals are clearly under the sur- face of the concept of explainability introduced before in this section.

To round up this prior analysis on the concept of explainability, the last

address explainability in ML models.

pretable by design, and those that can be explained by means of exter-

nal XAI techniques. This duality could also be regarded as the diﬀerence

post-hoc explainability. This same duality also appears in the paper pre-

methods to solve the transparent box design problem against the prob-

lem of explaining the black-box problem. This work, further extends

rules along with their speciﬁc output to aid in the understanding process

terms of the domain in which they are interpretable, namely, algorith-

next in connection to Fig. 3 , each of these classes contains its prede-

dominant place in this class. This being said, simple but extensive

within. This aspect aligns with the claim that sparse linear models

means of text and visualizations [32] . Again, endowing a decompos-

self-contained enough for a human to think and reason about it as a

•Decomposability stands for the ability to explain each of the parts of a

ability to understand, interpret or explain the behavior of a model.

premise). The added constraint for an algorithmically transparent

tic optimization (e.g. through stochastic gradient descent). The main

constraint for algorithmically transparent models is that the model

2.5.2. Post-hoc explainability techniques for machine learning models

Post-hoc explainability targets models that are not readily inter-

of the most common ways humans explain systems and processes by

nique. Overall, post-hoc explainability techniques are divided ﬁrst by

•Text explanations deal with the problem of bringing explainability for

explaining the results from the model [169] . Text explanations also

tioning of the model. These symbols may portrait the rationale of

•Visual explanation techniques for post-hoc explainability aim at vi-

in the model to users not acquainted to ML modeling. •Local explanations tackle explainability by segmenting the solution

these only explain part of the whole system’s functioning.

relate to the result generated by a certain model, enabling to get a

behave when attempting to explain a given process, explanations

by example are mainly centered in extracting representative exam-

which a whole new system is rebuilt based on the trained model to be

explained. This new, simpliﬁed model usually attempts at optimiz-

•Finally, feature relevance explanation methods for post-hoc explain-

an indirect method to explain a model.

when reviewing speciﬁc/agnostic XAI techniques for ML models in the

that maintains the model under the umbrella of transparent methods.

However, as stated in Section 2 , explainability is linked to a certain

post-hoc explainability techniques (mainly, visualization), particularly

when the model is to be explained to non-expert audiences.

of explaining the results of the models to non-expert users. Most authors

post-hoc explainability approaches available for a

Overall picture of the classiﬁcation of ML models attending to their level of explainability.

obtain the prediction of a

explain the knowledge

matches human naive

constrained within human

adopted for explaining these regression models. Visualization tech-

a model’s interpretation might entail, even when its construction is as

ear regression to maintain decomposability and simulatability, its size

ﬁll every constraint for transparency. Decision trees are hierarchical

fall out of the ﬁelds of computation and AI (even information technolo-

portance. Tree ensembles aim at overcoming such a poor performance

ent subsets of training data. Unfortunately, the combination of deci-

of post-hoc explainability techniques as the ones reviewed later in the

In terms of model explainability, it is important to observe that pre-

similarity between examples, which can be tailored depending on the

being simple to explain, the ability to inspect the reasons by which

over imprecise domains. Fuzzy systems improve two main axis relevant

rule systems in contexts with certain degrees of uncertainty. Rule based

plain complex models by generating rules that explain their predictions

intepretability. Similarly, the speciﬁcity of the rules plays also against

classical rules to fuzzy rules is to relax the constraints of rule sizes, since

makes them very suitable to understand and explain other models. If a

certain threshold of coverage is acquired, a rule wrapper can be thought

to contain enough information about a model to explain its behavior to a

one common factor: understandability. The main driver for conducting

in data. This is why GAMs have been accepted in certain communities

Once again, Bayesian models fall below the ceiling of Transparent

algorithmically transparent . However, it is worth noting that under certain

to explain other models, such as averaging tree ensembles [208] .

4. Post-hoc explainability techniques for machile learning

models: Taxonomy, shallow models and deep learning

to the model to explain its decisions. This is the purpose of post-hoc

explainability techniques (also referred to as post-modeling explainabil-

ity), which aim at communicating understandable information about

proaches for post-hoc explainability, discriminating among 1) those that

tiﬁed around post-hoc explainability for diﬀerent ML models, which are

•Model-agnostic techniques for post-hoc explainability ( Section 4.1 ),

•Post-hoc explainability that are tailored or speciﬁcally designed to

explain certain ML models. We divide our literature analysis into

two main branches: contributions dealing with post-hoc explain-

units ( Section 4.2 ); and techniques devised for deep learning models,

nation of Deep Learning models. To this end we focus on partic-

4.1. Model-agnostic techniques for post-hoc explainability

Model-agnostic techniques for post-hoc explainability are designed

times, simpliﬁed models are only representative of certain sections

the predictions of an opaque model to explain it. These contributions

G-REX has been extended to also account for model explainability

main ideas are exposed: a method for model distillation and com-

to check if the auditing data is missing key features it was trained

porally coincides with the most recent literature on XAI, including

this post-hoc explainability approach is envisaged to continue play-

ing a central role on XAI.

•Feature relevance explanation techniques aim to describe the function-

model to be explained. An amalgam of propositions are found within

Fig. 6. Taxonomy of the reviewed literature and trends identiﬁed for explainability techniques related to diﬀerent ML models. References boxed in blue, green

and red correspond to XAI techniques using image, text or tabular data, respectively. In order to build this taxonomy, the literature has been analyzed in depth to

discriminate whether a post-hoc technique can be seamlessly applied to any ML model, even if, e.g., explicitly mentions Deep Learning in its title and/or abstract.

of a classiﬁer trained with this subset of features cannot be distin-

model’s prediction back to the training data, by only requiring an

to its explainability [236,237] . Compared to those attempting ex-

found tackling explainability by means of feature relevance tech-

relevance has also become a vibrant subject study in the current XAI

mon in the ﬁeld of model-agnostic techniques for post-hoc explain-

rule extraction techniques prevail in model-agnostic contributions un-

der the umbrella of post-hoc explainability. This could have been intu-

explainability wrappers anticipated in Section 3.4 , and the complexity

dence of the feature relevance method on the model being explained.

the trained model (e.g. its structure, operations, etc) are tightly linked

to the speciﬁc model to be explained.

4.2. Post-hoc explainability in shallow ML models

Machines, SVMs) that require the adoption of post-hoc explainability

techniques for explaining their decisions.

combine diﬀerent trees to obtain an aggregated prediction/regression.

While it results to be eﬀective against overﬁtting, the combination of

from post-hoc explainability techniques. For tree ensembles, techniques

tree ensembles while maintaining part of the accuracy accounted for

the added complexity. The author from [119] poses the idea of training

not that many techniques to board explainability in tree ensembles by

crease Error) of the forest when a certain variable is randomly permuted

Finally, a crosswise technique among post-hoc explainability, [240] pro-

ensembles again, simpliﬁcation and feature relevance techniques seem

strategies, scarce activity has been recently noted around the explain-

interesting studies on the explainability of ensemble techniques include

much opaquer structure. Many implementations of post-hoc explainabil-

gin) to the nearest training-data point of any class, since in general, the

Section 2 , post-hoc explainability applied to SVMs covers explanation

trained model. This is the approach of [93] , which proposes a method

that extracts rules directly from the support vectors of a trained SVM us-

tors of a trained model. The work in [94] generates fuzzy rules instead

of authors considered adding the actual training data as a component

formulate the rule extraction problem as a multi-constrained optimiza-

plaining SVM models when used for concrete applications. For instance,

[77] presents an innovative approach to visualize trained SVM to extract

weights of a trained linear SVM that allows for a much more compre-

is speciﬁc enough to explain the multivariate patterns shown in neu-

tems, the latter being adopted as a post-hoc technique to explain de-

hoc explainability. In SVMs, local explanations have started to take some

plainability are dated beyond 2017, which might be due to the progres-

4.3. Explainability in deep learning

creasingly the most adopted methods for explaining DNNs. This section

reviews explainability studies proposed for the most used DL models,

their questionable explainability a common reason for reluctance. That

els. The fact that explainability is often a must for the model to be of

practical value, forced the community to generate multiple explainabil-

Some other works use model simpliﬁcation as a post-hoc explainabil-

is more complex as the number of layers increases, explaining these

of current explainability methods. For example, the authors in [262] ,

scores. This structure entails extremely complex internal relations that

are very diﬃcult to explain. Fortunately, the road to explainability for

skills favors the understanding of visual data. Existing works that aim at understanding what CNNs learn can be

curacy on several image recognition benchmarks. They obtained a

each ﬁlter to learn very speciﬁc object components. The obtained acti-

Fig. 7. Examples of rendering for diﬀerent XAI visualization techniques on images.

of a certain network [293] .

sented a three-level attention model to perform a ﬁne-grained classiﬁ-

ﬁlters out non-relevant patches to a certain object, and the last attention

der to analyse the visual information contained inside the CNN, the

the CNN internal representations and showed that several layers retain

paper also examines the eﬀects of classical training techniques on the

are arguably the most adopted approach to explainability in CNNs. Instead of using one single interpretability technique, the framework

A completely diﬀerent explainability approach is proposed in adver-

sarial detection. To understand model failures in detecting adversarial

from the representations of the training images.

As occurs with CNNs in the visual domain, RNNs have lately been

network considered as a black-box can be explained by associ-

Few contributions have been made for explaining RNN models.

These studies can be divided into two groups: 1) Explainability by un-

derstanding what a RNN model has learned (mainly via feature relevance

methods); and 2) explainability by modifying RNN architectures to pro-

dient Boosting Trees to the trained LSTM network under focus.

of the RNNs, [285] presents RETAIN (REverse Time AttentIoN) model,

or constraints in Knowledge Bases (KBs) has shown to not only improve

explainability but also performance with respect to purely data-driven

present in the training data labels. Other approaches have shown to be

models with explainability by externalizing other domain information

DNN, that must be supported by the training data [264] .

A diﬀerent perspective on hybrid XAI models consists of enriching

can be done by constraining the neural network thanks to a semantic

explainable autoencoders are proposed [313] . A speciﬁc transformer

relevant sections. Transformers can also help explain image captions vi-

the CBR (in this case a kNN) are paired in order to improve interpretabil-

CBR, in order to retrieve nearest-neighbor cases to explain the DNN’s

4.4. Alternative taxonomy of post-hoc explainability techniques for deep

erature on XAI. While the division between model-agnostic and model-

only relied on this criteria to classify XAI methods. For instance, some

model-agnostic methods such as SHAP [224] are widely used to explain

DL models. That is why several XAI methods can be easily categorized

Explaining a Deep Network Processing , as a kind of Linear Proxy Model .

plaining the processing of data by the network, thus answering to the

The second one concerns methods explaining the representation of data

the network contain? ”. The third approach concerns models speciﬁcally

structing XAI taxonomies.

Fig. 11. (a) Alternative Deep Learning speciﬁc taxonomy extended from the categorization from [13] ; and (b) its connection to the taxonomy in Fig. 6 . See References

Fig. 11 shows the alternative Deep Learning taxonomy inferred from

structure. This means that depending of the failure reasons of a complex

model, it would be possible to pick-up the right XAI method according

5. XAI: Opportunities, challenges and future research needs

addressed in the ﬁeld of explainability of ML and data fusion models.

them and explore new research opportunities for XAI, identifying pos-

Section 5.1 we will stress on the potential of XAI developments to

sensus on what explainability entails within the AI realm. Reasons

for pursuing explainability are also assorted and, under our own

•Given its notable prevalence in the XAI literature, Section 4.3 and

4.4 revolved on the explainability of Deep Learning models, examin-

several challenges that hold in regards to the explainability of this

connection to model explainability, remain insuﬃciently studied by

which enumerates research needs and open questions related to XAI

within a broader context: the need for responsible AI.

potential of XAI techniques and tools resides.

strained within very controlled physical problems, in which all of the

entails certain complexity, that the data available for study is greatly

the added complexity of the model will only ﬁght against the task of

tion of more sophisticated methods for explainability could invert or

spired by previous works [7] , in which XAI shows its power to improve

The literature clearly asks for an uniﬁed concept of explainability. In

the ﬁeld. It should propose a common structure for every XAI system.

This paper attempted a new proposition of a concept of explainability that is built upon that from Gunning [7] . In that proposition and the

following strokes to complete it ( Section 2.2 ), explainability is deﬁned

common ground and reference point to sustain a proﬁtable discussion in

this matter. It is paramount that the ﬁeld of XAI reaches an agreement in

Another key feature needed to relate a certain model to this con-

deﬁnition of explainable. Without such tool, any claim in this respect

should express how well the model performs in a certain aspect of ex-

plainability. Some attempts have been done recently around the mea-

surement of XAI, as reviewed thoroughly in [349,350] . In general, XAI

tal models, computational measures for explainer ﬁdelity, explanation

rection of evaluating XAI techniques. Unfortunately, conclusions drawn

quantiﬁable, general XAI metrics are really needed to support the exist-

whole prior acceptance of the broader concept of explainability, which

on the other hand, is one of the aims of the current work. Nevertheless,

performance of XAI techniques, as well as comparison methodologies

among XAI approaches that allow contrasting them quantitatively under

5.3. Challenges to achieve explainable deep learning

While many eﬀorts are currently being made in the area of XAI, there

are still many challenges to be faced before being able to obtain explain-

ability in DL models. First, as explained in Section 2.2 , there is a lack of

XAI. As an example, we often see the terms feature importance and feature

vations, attribution, and other approaches alike. As XAI is a relatively

why, as mentioned above, a challenge in XAI is establishing objective

explainable AI model are highlighted in [12] : First, explanations are

plained that probabilities are not as important as causal links in order

to translate the probabilistic results into qualitative notions contain-

meaning that focusing solely on the main causes of a decision-making

eﬃcient, while they oﬀer a greater explainability thus respecting the

•Being selective is less straightforward for connectionist models than

As stated in [24] , a truly explainable model should not leave expla-

A ﬁnal challenge XAI methods for DL need to address is provid-

ties, and to develop the social right to the (not-yet available) right

5.4. Explanations for AI security: XAI and adversarial machine learning

Nothing has been said about conﬁdentiality concerns linked to XAI.

thing secret , in the AI context many aspects involved in a model may

to make DL models more robust against intellectual property exposure

XAI tools capable of explaining ML models while keeping the model’s

Ideally, XAI should be able to explain the knowledge within an AI

However, the information revealed by XAI techniques can be used both

to generate more eﬀective attacks in adversarial contexts aimed at con-

protect against private content exposure by using such information. Ad-

[359] . For the particular case of DL models, available solutions such as

diﬀerent approaches to harden the model against them. Other examples

evasion attacks. There are even available solutions for unsupervised ML,

While XAI techniques can be used to furnish more eﬀective adversar-

other generative models towards explaining data-based decisions. Once

trained, generative models can generate instances of what they have

adopted by several recent studies [366,367] mainly as an attribution

method to relate a particular output of a Deep Learning model to their

5.5. XAI and output conﬁdence

pend on the output of AI models, such as vehicular perception and self-

yielded comprehensive regulatory eﬀorts aimed at ensuring that no de-

risk and uncertainty of harms derived from decisions made on the out-

the share of epistemic uncertainty (namely, the uncertainty due to lack

jection of the model’s output [370,371] . To this end, explaining via XAI

uncertainty within the input domain.

5.6. XAI, Rationale explanation, and critical data studies

also to the availability of information about the full discourse around

this discourse we ﬁnd also an interesting space for the adoption of XAI

XAI can eﬀectively ease the process of explaining the reasons why a

make them appraise the ethical implications of their data-based choices

where XAI can signiﬁcantly boost the exchange of information among

5.7. XAI And theory-guided data science

We envision an exciting synergy between the XAI realm and Theory-

is previously known. Similarly, the training approach should not allow

available with promising results. The studies in [375–382] were carried

XAI. All the additions presented in [374] push toward techniques that

would eventually render a model explainable, and furthermore, knowl-

tured by a model should be explained for assessing its compliance with

theoretical principles known beforehand. This, again, opens a magniﬁ-

cent window of opportunity for XAI.

5.8. Guidelines for ensuring interpretable AI models

nature of the process of making an AI-based model interpretable. Along

teracting with the system to be explained, from the designers of the sys-

guidelines to implement and explain AI systems have been recently con-

suggests that the incorporation and consideration of explainability in

practical AI design and deployment workﬂows should comprise four ma-

1. Contextual factors, potential impacts and domain-speciﬁc needs

for which the AI model is built, the complexity of explanations that

a reference point for the AI system to be deployed in lieu thereof.

considering explainability in the development of an AI system, the

decision of which XAI approach should be chosen should gauge

domain-speciﬁc risks and needs, the available data resources and

existing domain knowledge, and the suitability of the ML model to

and domain-speciﬁc needs) can make transparent models preferable

the application of post-hoc XAI techniques. By contrast, black-box

that ethics-, fairness- and safety-related impacts should be weighed.

AI system should be ensured by checking whether such identiﬁed

system with XAI tools that provide the level of explainability re-

quired by the domain in which it is deployed. To this end, the third

guideline suggests 1) a detailed articulation, examination and eval-

whether the coverage and scope of the available explanatory ap-

proaches match the requirements of the domain and application

livery strategy, including a detailed time frame for the execution of

measures of explainability are intensively revolving by considering

AI. Methodological principles ensure that the purpose for which explain-

relevance such as no discrimination, sustainability, privacy or account-

ability. A challenge remains in harnessing the potential of XAI to realize

a Responsible AI , as we discuss in the next section.

6. Toward responsible AI: Principles of artiﬁcial intelligence,

fairness, privacy and data fusion

published guidelines to indicate how AI should be developed and used.

These guidelines are commonly referred to as AI principles , and they

tackle issues related to potential AI threats to both individuals and to

and widely recognized principles in order to link XAI –which normally

implementation and use of AI models be sought in practice, it is our ﬁrm

claim that XAI does not suﬃce on its own. Other important principles of

Artiﬁcial Intelligence such as privacy and fairness must be carefully ad-

of Responsible AI, along with the implications of XAI and data fusion in

A recent review of some of the main AI principles published since

explainability, or fairness. They also consider the coverage that the

•Target audience: to whom the principles are aimed. They are nor-

For instance, [386] is an illustrative example of a document of AI

of the most common principles, and deals explicitly with explainability.

Here, the authors propose ﬁve principles mainly to guide the develop-

ment of AI within their company, while also indicating that they could

The authors of those principles aim to develop AI in a way that

•The outputs after using AI systems should not lead to any kind of

discrimination against individuals or collectives in relation to race,

while optimizing the results of an AI system is not only their outputs

those groups. This deﬁnes the principle of Fair AI .

person, and when they are communicating with an AI system. Peo-

by the AI system and for what purpose. It is crucial to ensure a

certain level of understanding about the decisions taken by an AI

system. This can be achieved through the usage of XAI techniques.

Transparent and Explainable AI .

•AI products and services should always be aligned with the United

Nation’s Sustainable Development Goals [387] and contribute to

them in a positive and tangible way. Thus, AI should always gen-

principle of Human-centric AI (also referred to as AI for Social Good

•AI systems, specially when they are fed by data, should always con-

principle is not exclusive of AI systems since it is shared with many

and recommendations aimed at considering a wider context for sci-

challenges such as sustainability, public engagement, ethics, science

ple of Transparent and Explainable AI mentioned previously.

Going beyond the scope of these ﬁve AI principles, the European

worthy AI [390] through an assessment checklist that can be completed

by diﬀerent proﬁles related to AI systems (namely, product managers,

discrimination and fairness; 5) societal and environmental well-being;

6) accountability. These principles are aligned with the ones detailed

including any type of organization involved in the development of AI.

It is worth mentioning that most of these AI principles guides directly

approach XAI as a key aspect to consider and include in AI systems. In

that 28 out of the 32 AI principles guides covered in the analysis, explic-

itly include XAI as a crucial component. Thus, the work and scope of this

AI at a worldwide level.

6.2. Fairness and accountability

pects, beyond XAI, included within the diﬀerent AI principles guidelines

pletely detached from XAI; in fact, they are intertwined. This section

presents two key components with a huge relevance within the AI prin-

ciples guides, Fairness and Accountability. It also highlights how they

are connected to XAI.

6.2.1. Fairness and discrimination

tected and unprotected features where XAI techniques ﬁnd their place

ables amenable to cause discrimination. XAI techniques such as SHAP

[224] could be used to generate counterfactual outcomes explaining the

Recalling the Fair AI principle introduced in the previous section,

[386] reminds that fairness is a discipline that generally includes pro-

unintentionally create unfair decisions by considering sensitive factors

unfair decisions can give rise to discriminatory issues, either by explic-

aforementioned proposals centered on fairness aspects permit to dis-

•Individual fairness: here, fairness is analyzed by modeling the diﬀer-

•Group fairness: it deals with fairness from the perspective of all in-

•Counterfactual fairness: it tries to interpret the causes of bias using,

•Tainted data: Errors in the data modelling deﬁnition, wrong feature

to deﬁne when AI is not biased. For supervised ML, [393] presents a

framework that uses three criteria to evaluate group fairness when there

fairness.

ML model is trained, looking to remove the bias at the ﬁrst step of

•In-processing: These techniques are applied during the training pro-

cess of the ML model. Normally, they include Fairness optimization

constraints along with cost functions of the ML model. An example

is trained. They are less intrusive because they do not modify the

Even though these references apparently address an AI principle that

appears to be independent of XAI, the literature shows that they are in-

AI principles that deal with XAI, also talk about fairness explicitly. This

when implementing Responsible AI.

The literature also exploses that XAI proposals can be used for bias

the bias present in a model (both for individual and group fairness).

Thus, the fairness report is shown just like the visual summaries used

within XAI. This explainability approach eases the understanding and

it quantitatively, indicate the degree of fairness, and explain why a user

or group would be treated unfairly with the available data. Similarly,

XAI techniques such as SHAP [224] could be used to generate coun-

terfactual outcomes explaining the decisions of a ML model when fed

lations between protected and unprotected features through XAI tech-

Another example is [399] , where the authors propose a fair-by-

a small part of the whole dataset available (weak supervision). It ﬁrst

it generates rules in an IF/THEN format that explain that a data point

cently pursued in [401] , showing that post-hoc XAI techniques can forge

fairer explanations from truly unfair black-box models. Finally, CERTI-

FAI (Counterfactual Explanations for Robustness, Transparency, Inter-

pretability, and Fairness of Artiﬁcial Intelligence models) [402] uses a

ine fairness (both at the individual level and at the group level) at the

Strongly linked to the concept of fairness, much attention has been

with ethical restrictions that permeate to the AI modeling phase [404] .

Likewise, certain AI problems (such as content recommendation or infor-

mation retrieval) also aim at producing diverse recommendations rather

secting the internals of a black-box model via XAI techniques can help

identifying the capability of the model to maintain the input data di-

sity keeping capabilities could be complemented with XAI techniques

from which the model was trained. Conversely, XAI could help to dis-

AI systems. Performing the assessment by both internal and external

auditors, and making the reports available, could contribute to the

trustworthiness of the technology. When the AI system aﬀects funda-

porting actions or decisions that yield a certain outcome by the sys-

to respond to them. To address that, the development of AI systems

and use of AI systems. It is also important to guarantee protection

for anyone who raises concerns about an AI system (e.g., whistle-

AI systems pose. •Trade-oﬀs: In case any tension arises due to the implementation of

use of the AI system should not proceed in that form.

sure trust. Special attention should be paid to vulnerable persons or

of XAI with accountability. First, XAI contributes to auditability as it

can help explaining AI systems for diﬀerent proﬁles, including regula-

tory ones. Also, since there is a connection between fairness and XAI as

stated before, XAI can also contribute to the minimization and report of

exist in almost all domains of activity calls for data fusion approaches

aimed at exploiting them simultaneously toward solving a learning task.

the potential of data fusion techniques to enrich the explainability of

sible AI, the conﬂuence between XAI and data fusion is an uncharted

research area in the current research mainstream.

( Fig. 13 .e). Upon local model training, clients transmit encrypted in-

object as per the information contained in the diﬀerent data sources

to diﬀerent algorithmic means, from co-training to co-regularization

responsible AI paradigm

AI systems, specially when dealing with multiple data sources, need

•Privacy and data protection: they should be guaranteed in AI systems

user could be used in a negative way against them (discrimination

due to sensitive features, unfair treatment...), it is crucial to ensure

reach good performance with AI systems that are fueled with data,

like ML. However, sometimes the data collected contains socially

be tackled before training any model with the data collected. Addi-

is directly intertwined with privacy and with fairness, regardless of the

ing input queries on the model [356,357] . An approach to explain loss

direction, namely, in ensuring that XAI methods do not pose a threat in

regards to the privacy of the data used for training the ML model under

the context of explainability covered in this survey. To begin with, clas-

no connection to the ML model, so they have little to do with explain-

can be thought to aim at solving a data level fusion problem, yet in a

In this context, many techniques in the ﬁeld of XAI have been pro-

paves the way to explaining how data sources are actually fused through

spatial and/or time domain. Ultimately, this gained information on the

level contemplates data under certain constraints of known form and

in the possibility that XAI techniques could be explanatory enough to

the explained fusion among protected and unprotected features.

making it necessary to resort to post-hoc explainability solutions. How-

ever, model fusion may entail other drawbacks when endowed with

powerful post-hoc XAI techniques. Let us imagine that relationships of

tend to) models are fused. These models contain among others, cell-

If focused at knowledge level fusion, a similar reasoning holds: XAI

ability to explain models could have an impact on the necessity of

within ML models. If so, XAI might enrich knowledge fusion paradigms,

portance that the knowledge extracted from a model by means of XAI

techniques can be understood and extrapolated to the domain in which

of transfer learning portrayed in [425] . Although XAI is not contem-

trained in certain feature spaces and distributions, to then be utilized in

XAI can pose a threat if the explanations given about the model can be

spurs further challenges in regards to privacy and explainability. The

Distributed fusion might be applied for diﬀerent reasons, mainly due to

environmental constraints or due to security or privacy issues. The latter

of an ML model trained on local data. This rationale lies at the heart

This lightens the training process for network-compromised settings and

guarantees data privacy [416] . Upon the use of post-hoc explainability

cal context in which the received ML model part was trained. In fact,

every single hyperparameter value used for training, allowing for poten-

Data fusion, privacy and model explainability are concepts that have

6.4. Implementing responsible AI principles in an organization

While increasingly more organizations are publishing AI principles

Fig. 14. Summary of XAI challenges discussed in this overview and its impact on the principles for Responsible AI.

•AI-speciﬁc principles that focus on aspects that are speciﬁc to AI,

such as explainability, fairness and human agency.

•End-to-end principles that cover all aspects involved in AI, including

The EC Guidelines for Trustworthy AI are an example of end-to-end

operating worldwide) are more AI-speciﬁc [386] . For example, safety

also for AI systems. The same holds for privacy, but it is probably true

that privacy in the context of AI systems is even more important than for

data and most importantly, because XAI tools and data fusion techniques

When it comes to implement the AI Principles into an organization, it

is important to operationalize the AI-speciﬁc parts and, at the same time,

for privacy, security and safety. Implementing AI principles requires a

•AI principles (already discussed earlier), which set the values and

•Awareness and training about the potential issues, both technical

•A questionnaire that forces people to think about certain impacts of

the AI system ( impact explanation ). This questionnaire should give

concrete guidance on what to do if certain undesired impacts are

ing any problems identiﬁed. XAI tools and fairness tools fall in this

nance: 1) based on committees that review and approve AI devel-

Responsible AI principles in companies should balance between two re- quirements: 1) Major cultural and organizational changes needed to en-

force such principles over processes endowed with AI functionalities;

principles with the IT assets, policies and resources already available at

around the principles and values of Responsible AI where we envision

that XAI will make its place and create huge impact.

This overview has revolved around eXplainable Artiﬁcial Intelli-

gence (XAI), which has been identiﬁed in recent times as an utmost need

model explainability, as well as by showing the diverse purposes that

cent literature dealing with explainability, which has been approached

and 2) post-hoc XAI techniques devised to make ML models more in-

explainability of Deep Learning models, we have inspected in depth the

tive taxonomy that connects more closely with the speciﬁc domains in

which explainability can be realized for Deep Learning models.

in the XAI realm toward the concept of Responsible AI, a paradigm that

imposes a series of AI principles to be met when implementing AI models

in practice, including fairness, transparency, and privacy. We have also

discussed the implications of adopting XAI techniques in the context of

data fusion, unveiling the potential of XAI to compromise the privacy

of protected data involved in the fusion process. Implications of XAI in

fairness have also been discussed in detail. This vision of XAI as a core

concept to ensure the aforementioned principles for Responsible AI is

Our reﬂections about the future of XAI, conveyed in the discus-

XAI techniques. It is our vision that model interpretability must be ad-

dressed jointly with requirements and constraints related to data pri-

vacy, model conﬁdentiality, fairness and accountability. A responsible

implementation and use of AI methods in organizations and institutions

worldwide will be only guaranteed if all these AI principles are studied

received through the EMAITEK and ELKARTEK programs. Javier Del

research and innovation programme AI4EU under grant agreement

ble AI: R. Benjamins, A. Barbado, D. Sierra, “Responsible AI by Design ”, to

appear in the Proceedings of the Human-Centered AI: Trustworthiness

of AI Models & Data (HAI) track at AAAI Fall Symposium, DC, November

[2] D.M. West , The future of work: robots, AI, and automation, Brookings Institution

ing and a “right to explanation ”, AI Magazine 38 (3) (2017) 50–57 .

[4] D. Castelvecchi , Can we open the black box of AI? Nature News 538 (7623) (2016)

[6] A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in

Explainable AI, 2018.

[7] D. Gunning , Explainable artiﬁcial intelligence (xAI), Technical Report, Defense Ad-

[8] E. Tjoa, C. Guan, A survey on explainable artiﬁcial intelligence (XAI): Towards

medical XAI, 2019.

[9] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G.M. Youngblood, Explainable AI for

[10] F.K. Dos ̃ilovi ć, M. Brc ̃i ć, N. Hlupi ć, Explainable artiﬁcial intelligence: A survey,

[13] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining Explana-

[14] A. Adadi , M. Berrada , Peeking inside the black-box: A survey on explainable arti-

ﬁcial intelligence (XAI), IEEE Access 6 (2018) 52138–52160 .

in: IJCAI-17 workshop on explainable AI (XAI), 8, 2017, p. 1 .

man-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications

and Bibliography for Explainable AI, Technical Report, Defense Advanced Research

Projects Agency (DARPA) XAI Program, 2019 .

of methods for explaining black box models, ACM Computing Surveys 51 (5) (2018)

ary fuzzy systems for explainable artiﬁcial intelligence: Why, when, what for, and

[21] M.W. Craven , Extracting comprehensible models from trained neural networks,

[24] D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? a new

[29] F. Rossi, AI Ethics for Enterprise AI, 2019.

plainable Ai systems for the medical domain?, 2017.

SAIL-TR-2015-010, 2015 .

[32] M.T. Ribeiro , S. Singh , C. Guestrin , Why should I trust you?: Explaining the pre-

[33] M. Fox, D. Long, D. Magazzeni, Explainable planning, 2017.

[34] H.C. Lane , M.G. Core , M. Van Lent , S. Solomon , D. Gomboc , Explainable artiﬁcial

intelligence for training and tutoring, Technical Report, University of Southern

[37] A. Chander , R. Srinivasan , S. Chelian , J. Wang , K. Uchino , Working with beliefs: AI

rections and challenges in extracting the knowledge embedded within trained arti-

[40] O. Goudet , D. Kalainathan , P. Caillou , I. Guyon , D. Lopez-Paz , M. Sebag , Learn-

ing functional causal models with generative neural networks, in: Explainable and

[43] C. Barabas, K. Dinakar, J. Ito, M. Virza, J. Zittrain, Interventions over predictions:

[46] W. Samek, T. Wiegand, K.-R. Müller, Explainable artiﬁcial intelligence: Under-

standing, visualizing and interpreting deep learning models, 2017.

[47] C. Wadsworth, F. Vera, C. Piech, Achieving fairness through adversarial learning:

deep learning, IEEE Transactions on Neural Networks and Learning Systems 30

[50] M. Harbers , K. van den Bosch , J.-J. Meyer , Design and evaluation of explainable

son , T.C. Bailey , A. Hernandez , Conﬁdent interpretation of bayesian decision tree

[58] F.J.C. Garcia, D.A. Robb, X. Liu, A. Laskov, P. Patron, H. Hastie, Explain yourself:

[59] P. Langley , B. Meadows , M. Sridharan , D. Choi , Explainable agency for intelli-

gent autonomous systems, in: AAAI Conference on Artiﬁcial Intelligence, 2017,

[60] G. Montavon , S. Lapuschkin , A. Binder , W. Samek , K.-R. Müller , Explaining non-

Learning how to explain neural networks: Patternnet and patternattribution, 2017.

[62] G. Ras , M. van Gerven , P. Haselager , Explanation methods in deep learning: Users,

values, concerns and challenges, in: Explainable and Interpretable Models in Com-

autoencoders, in: Paciﬁc Symposium on Biocomputing Co-Chairs, World Scientiﬁc,

trained neural networks: a practical and eﬃcient approach, IEEE Transactions on

sity: Tree regularization of deep models for interpretability, in: AAAI Conference

[83] Z.-H. Zhou , Y. Jiang , S.-F. Chen , Extracting symbolic rules from trained neural

network ensembles, AI Communications 16 (1) (2003) 3–15 .

[86] T. Miller , P. Howe , L. Sonenberg , Explainable AI: Beware of inmates running the

Explainable AI (XAI), 36, 2017, pp. 36–40 .

berg , A. Holzinger , Explainable AI: the new 42? in: International Cross-Domain

[88] V. Belle , Logic meets probability: Towards explainable AI systems for uncer-

tain worlds, in: International Joint Conference on Artiﬁcial Intelligence, 2017,

[90] Y. Lou , R. Caruana , J. Gehrke , G. Hooker , Accurate intelligible models with pair-

[97] R. Krishnan , G. Sivakumar , P. Bhattacharya , Extracting decision trees from trained

[99] B. Green , “Fair ”r i s k assessments: A precarious approach for criminal justice re-

form, in: 5th Workshop on Fairness, Accountability, and Transparency in Machine

[100] A. Chouldechova , Fair prediction with disparate impact: A study of bias in recidi-

[101] M. Kim , O. Reingold , G. Rothblum , Fairness through computationally-bounded

[106] Y. Zhang , H. Su , T. Jia , J. Chu , Rule extraction from trained support vector ma-

models using transparent model distillation, in: AAAI/ACM Conference on AI,

[128] M. Kearns, S. Neel, A. Roth, Z.S. Wu, Preventing fairness gerrymandering: Auditing

and learning for subgroup fairness, 2017.

[135] J.J. Thiagarajan, B. Kailkhura, P. Sattigeri, K.N. Ramamurthy, Treeview: Peeking

pretable and ﬁne-grained visual explanations for convolutional neural networks,

[139] A. Kanehira , T. Harada , Learning to explain with complemental examples, in: Pro-

The all convolutional net, 2014. [144] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Inter-

trained neural networks, in: Machine learning proceedings 1994, Elsevier, 1994,

[148] A.D. Arbatli , H.L. Akin , Rule extraction from trained neural networks using genetic

[155] P.E. Pope , S. Kolouri , M. Rostami , C.E. Martin , H. Hoﬀmann , Explainability meth-

[156] P. Gajane, M. Pechenizkiy, On formalizing fairness in prediction with machine

[157] C. Dwork, C. Ilvento, Composition of fairsystems, 2018.

[158] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairml-

book.org, 2019. http://www.fairmlbook.org

[169] A. Bennetot , J.-L. Laurent , R. Chatila , N. Díaz-Rodríguez , Towards explainable neu-

ral-symbolic visual reasoning, in: NeSy Workshop IJCAI 2019, Macau, China, 2019 .

[172] K. Kawaguchi , Deep learning without poor local minima, in: Advances in neural

[183] L. Rokach , O.Z. Maimon , Data mining with decision trees: theory and applications,

opaque models using genetic programming., in: FLAIRS Conference, Miami Beach,

[191] J.R. Quinlan , Generating production rules from decision trees., in: ijcai, 87, Cite-

factors and distributions of pelagic ﬁsh and krill: a case study in sendai bay, Japan,

[207] A.R. Cassandra , L.P. Kaelbling , J.A. Kurien , Acting under uncertainty: Discrete

tions by identifying prediction invariance, 2016. [217] M.W. Craven , Extracting Comprehensible Models from Trained Neural Networks,

1996 Ph.D. thesis . AAI9700774

[226] H. Chen, S. Lundberg, S.-I. Lee, Explaining models by propagating shapley values

[229] J. Moeyersoms, B. d’Alessandro, F. Provost, D. Martens, Explaining classiﬁcation

How to explain individual classiﬁcation decisions, Journal of Machine Learning

[234] M. Robnik- Š ikonja , I. Kononenko , Explaining classiﬁcations for individual in-

planations, in: AAAI Conference on Artiﬁcial Intelligence, 2018, pp. 1527–

[236] D. Martens , F. Provost , Explaining data-driven document classiﬁcations, MIS Quar-

[237] D. Chen , S.P. Fraiberger , R. Moakler , F. Provost , Enhancing transparency and con-

[240] G. Tolomei , F. Silvestri , A. Haines , M. Lalmas , Interpretable predictions of

[243] N.F. Rajani , R.J. Mooney , Ensembling visual explanations, in: Explainable and In-

[256] H. Tsukimoto , Extracting rules from trained neural networks, IEEE Transactions on

[260] R. Féraud , F. Clérot , A methodology to explain neural network classiﬁcation, Neural

pretable and robust deep learning, 2018.

[271] Y. Goyal, A. Mohapatra, D. Parikh, D. Batra, Towards transparent AI systems: In-

attention models in deep convolutional neural network for ﬁne-grained image clas-

[280] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural net-

[282] J. Clos , N. Wiratunga , S. Massie , Towards explainable text classiﬁcation by jointly

learning lexicon and modiﬁer terms, in: IJCAI-17 Workshop on Explainable AI

(XAI), 2017, p. 19 .

[285] E. Choi , M.T. Bahadori , J. Sun , J. Kulas , A. Schuetz , W. Stewart , Retain: An inter- pretable predictive model for healthcare using reverse time attention mechanism,

[287] A. Lucic, H. Haned, M. de Rijke, Explaining predictions from tree-based boosting

[290] R. Traoré, H. Caselles-Dupré, T. Lesort, T. Sun, G. Cai, N.D. Rodríguez, D. Filliat,

works with applications to healthcare domain, 2015.

Artiﬁcial Intelligence, IJCAI (2017) 1596–1602 .

on dietary data, in: First Workshop on Semantic Explainability @ ISWC 2019, 2019 .

ence on Artiﬁcial Intelligence, IJCAI-18, 2018, pp. 1362–1368 .

[311] M. Zolotas , Y. Demiris , in: Towards explainable shared control using augmented

toencoders for explainable recommender systems, in: Proceedings of the 3rd Work-

shop on Deep Learning for Recommender Systems, in: DLRS 2018, 2018, pp. 24–31 .

[314] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A.M. Dai, M.D.

[315] M. Cornia, L. Baraldi, R. Cucchiara, Smart: Training shallow memory-aware trans-

formers for robotic explainability, 2019.

XAI: An Overview of ANN-CBR Twins for Explaining Deep Learning, 2019.

[319] T. Hailesilassie, Rule extraction algorithm for deep neural networks: A review,

[328] Y. Zhang, X. Chen, Explainable Recommendation: A Survey and New Perspectives,

[329] J. Frankle, M. Carbin, The Lottery Ticket Hypothesis: Finding Sparse, Trainable

[330] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser,

[334] A. Slavin Ross, M.C. Hughes, F. Doshi-Velez, Right for the Right Reasons: Training

Diﬀerentiable Models by Constraining their Explanations, 2017.

A. Lerchner , beta-vae: Learning basic visual concepts with a constrained variational

Networks to Explain Neural Networks, 2018.

[345] D. Bouchacourt, L. Denoyer, EDUCE: explaining model decisions through unsuper-

[347] C. Rudin, Please stop explaining black box models for high stakes decisions, 2018.

[349] R.R. Hoﬀman, S.T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Chal-

design and evaluation of explainable ai systems, 2018.

[351] R.M.J. Byrne , Counterfactuals in explainable artiﬁcial intelligence (XAI): Evidence

Conference on Artiﬁcial Intelligence, IJCAI-19, 2019, pp. 6276–6282 .

[352] M. Garnelo , M. Shanahan , Reconciling deep learning with symbolic artiﬁcial intel-

works, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,

Springer, 2019, pp. 121–144 . [358] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial ex-

Kohno, D. Song, Robust physical-world attacks on deep learning models, 2017.

[362] B. Biggio , I. Corona , D. Maiorca , B. Nelson , N. Š rndi ć, P. Laskov , G. Giacinto , F. Roli ,

Evasion attacks against machine learning at test time, in: Proceedings of the 2013th

[363] B. Biggio, I. Pillai, S.R. Bulò, D. Ariu, M. Pelillo, F. Roli, Is data clustering in adver-

[367] C. Biﬃ, O. Oktay , G. Tarroni , W. Bai , A. De Marvao , G. Doumou , M. Rajchl ,

R. Bedair , S. Prasad , S. Cook , et al. , Learning interpretable anatomical features

[368] S. Liu, B. Kailkhura, D. Loveland, Y. Han, Generative counterfactual introspection

for explainable deep learning, 2019.

[375] G. Hautier , C.C. Fischer , A. Jain , T. Mueller , G. Ceder , Finding nature’s missing

gap-ﬁlling and trait prediction for macroecology and functional biogeography,

[384] C. Rudin , Stop explaining black box machine learning models for high stakes deci-

[386] R. Benjamins, A. Barbado, D. Sierra, Responsible AI by design, 2019.

[387] United-Nations , Transforming our World: the 2030 Agenda for Sustainable Devel-

[389] B.C. Stahl , D. Wright , Ethics and privacy in ai and big data: Implementing respon-

thy AI, Technical Report, European Commission, 2019 .

M.B. Zafar , A uniﬁed approach to quantifying algorithmic unfairness: Measuring

individual group unfairness via inequality indices, in: Proceedings of the 24th ACM

[396] R. Zemel , Y. Wu , K. Swersky , T. Pitassi , C. Dwork , Learning fair representations,

learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and

[398] Y. Ahn , Y.-R. Lin , Fairsight: Visual analytics for fairness in decision making, IEEE

[399] E. Soares , P. Angelov , Fair-by-design explainable models for prediction of recidi-

[400] J. Dressel , H. Farid , The accuracy, fairness, and limits of predicting recidivism,

[401] U. Aivodji , H. Arai , O. Fortineau , S. Gambs , S. Hara , A. Tapp , Fairwashing: the

[402] S. Sharma , J. Henderson , J. Ghosh , Certifai: Counterfactual explanations for robust-

ness, transparency, interpretability, and fairness of artiﬁcial intelligence models,

ference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 220–229 .