EKONOMIA I PRAWO. ECONOMICS AND LAW
Volume 19, Issue 3, September 2020
p-ISSN 1898-2255, e-ISSN 2392-1625
www.economicsandlaw.pl

ORIGINAL ARTICLE
received 27.03.2020; revised 20.05.2020; accepted 30.09.2020
Citation: Bejger, S., & Elster, S. (2020). Artificial Intelligence in economic decision making: how to assure a trust?. *Ekonomia i Prawo. Economics and Law*, 19(3): 411–434.
doi:10.12775/EiP.2020.028.

Artificial Intelligence in economic decision making: how to assure a trust?

SYLWESTER BEJGER

corresponding author
Nicolaus Copernicus University in Toruń, Faculty of Economic Sciences and Management,
Department of Applied Informatics and Mathematics in Economics,
ul. Gagarina 13a, 87-100 Toruń, Poland
✉ sylw@umk.pl
orcid.org/0000-0001-7900-946X

STEPHAN ELSTER

Nicolaus Copernicus University in Toruń, Poland
✉ stephan.elster@doktorant.umk.pl
orcid.org/0000-0002-9896-8970

Abstract

Motivation: The decisions made by modern 'black box' artificial intelligence models are not understandable and therefore people do not trust them. This limits down the potential power of usage of Artificial Intelligence.

Aim: The idea of this text is to show the different initiatives in different countries how AI, especially black box AI, can be made transparent and trustworthy and what kind of regulations will be implemented or discussed to be implemented. We also show up how a commonly used development process within Machine Learning can be enriched to fulfil the requirements e.g. of the Ethics guidelines for trustworthy AI of the High-Level Expert Group of the European Union. We support our discussion with a proposition of empirical tools providing interpretability.

Results: The full potential of AI or products using AI can only be raised if the decision of AI models are transparent and trustworthy. Regulations which are followed over the whole life cycle of AI models, algorithms or the products they using these are therefore necessary as well as understandability or explainability of the decisions these models and algorithms made. Initiatives on every level of stakeholders started, e.g. international level on the European Union, country level, USA, China etc. as well on a company level.

The post-hoc local interpretability methods could and should be implemented by economic decision makers to provide compliance with the regulations.

Keywords: AI regulations; AI law; explainability of AI; local intepretability
JEL: C38; D83; D85; H10; G38

1. Introduction

If people can't understand the decisions of artificial intelligence, they don't trust them, whether it's for approving a loan application or in autonomous driving. This is equally true in a case of a provider of an Artificial Intelligence (AI) service as well as an end user of such service. Such 'understandability' is especially important in managerial decisions on various levels and in various organizations (private, public, governmental). The decisions, actions of the AI models and the models itself must therefore be transparent, explainable and interpretable. Such paradigm is enforced by regulations and informal standards in various countries and organizations (Algorithm Watch, 2020; Dutton, 2018; Library of Congress, 2019; Wischmeyer, 2020).

In the machine learning literature, early work on explanation focused on producing visualizations of the prediction (symbolic AI or glass-box AI). The results or decisions made could easily be understand by humans. The present most successful AI models and algorithms are based on statistical methods or statistical learning methods and the results decisions and actions, are not understandable to a human entity. Present works focused on two approaches to explanation: interpretation and justification of prediction of a specific model in a specific domain and the intrinsic interpretability of models (DARPA, 2016; de Laat, 2018; Holzinger, 2018; Molnar, 2020).

In our work we would like to underline an importance of interpretability of models. We understand it not only as an ability to predict what is going to happen with the output of the model, given a change in input or algorithmic parameters but as a business domain — linked understanding of influence of algorithmic parameters on modelling phenomena. We think that assessment of interpretability and explainability should be defined as part of a business process and integrated within the process map of an organization. We propose an adequate modification of Cross Industry Standard Process for Data Mining (CRISP DM) as a sub process as well. We point to formal methods of providing a trust for AI, so called post-hoc interpretability methods and evaluated one of them on a basis of economic decision use case.

2. Literature review

There is a growing interest in the topics of AI and ethics or robot-ethics. Authors equalize AI and robotics. We agree to that as the main component of an artificial intelligent robotic system is its AI that can be seen therefore as the basis for such a robotic system. We concentrate in the following on a part of AI methods like e.g. machine learning. Regulating of AI always needs a deep look into ethics. So, some authors discuss how to develop a friendly AI which is benevolent to human society (Anderson, & Anderson, 2007; Yudkowsky, 2001). There is a growing interest in pessimistic thinking about how robots or AI will change the society and cause the loss of many jobs (Brynjolfsson & McAfee, 2011; Ford, 2015). On the other side there are more optimistic views on AI from e.g. Kurzweil (2005) and we have also more pessimistic views on AI of e.g. Bostrom (2014). A lot of new literature is discussing the need for an interpretable and transparent AI, therefore many Authors are demanding an interpretability (explainability) and transparency of AI decisions as well as accountability on all levels of society as well as over the whole lifecycle of the models and algorithms of AI. Doing so the authors discuss how a regulation of AI (or robots) can look like (Gunkel, 2018; Turner, 2019). Authors who are focusing on the decisions and transparency of AI models describe the change from the glass-box models using human understandable symbolic results in the early stage of AI research towards the black-box models using statistical algorithms, e.g. partially for a black box model or parts of it a better and easier explainable model, or the vary the input variables and parameter and try to explain the different results. Many authors demand also accountability and differentiate between the model/ algorithm or the product using such, e.g. a robot, and the stakeholders of AI on different levels and different phases of the AI model lifecycle (Sauerwein, 2019; Wischmeyer, 2020). A lot of countries are discussing regulations for AI or products using AI (Algorithm Watch, 2020; Dutton, 2018; Library of Congress, 2019).

3. Methods

When writing the article, the authors used the methods of inductive and deductive reasoning as well as descriptive analysis and enrichment of the commonly used CRISP–DM process. The author also used the method of comparative analysis. The research process involved the identification of a research problem, that is the problem of lack of regulations and legislation for AI and the lack of trustworthiness and non explainability of decisions and actions made by AI algorithms and models. Next we pointed to a set of formal (technical) methods of interpretability and examined one of them, doing empirical research with data. The findings of the research of the started and ongoing initiatives and the idea of implementation of regulations lead to an example how these can be.

4. Results

Only if people can understand and trust the decisions of AI the full potential of the models can be utilized, e.g. by using stochastic methods like artificial neural networks within machine learning (ML) to imitate human intelligence and doing so providing neutral recommendations, decisions and actions with analysing more volumes of data faster than any human being. As there are more and more decisions done by AI — also in some critical domains, like justice, loan approval, selection of personnel, medicine, traffic or military, transparency is a basis, a 'condition sine qua non', to trust the decisions of AI. There is therefore a high demand of regulations how to provide such a transparency on different levels of organizations, e.g. on European Union level or international, national or inner-organizational and also in different phases of the lifecycle of such models (DARPA, 2016; de Laat, 2018; Holzinger, 2018; Waltl & Vogl, 2018).

Understand ability and transparency are the basis for accountability, therefore many authors demand an accountability on different levels, like society, state or enterprise level (e.g. Sauerwein, 2019). Also there are different phases in the lifecycle of AI models and there are different stakeholders who influence the model at different stages in its lifecycle. At a technology level, an 'accountability by design' must be implemented, in which the designers commit themselves, for example, to Responsible Research and Innovation (RRI) like a Hippocratic Oath for developer. Another possible way might be the use of appropriate methods, e.g. by enrichment of the Cross Industry Standard Process for Data Mining (CRISP–DM) process and the requirements demanded by ethical guidelines or regulations. Companies that use AI in their products need to document their social responsibility, for example, by means of Corporate Social Responsibility (CSR). By following given ethical guidelines also in the development process, they might get a certification by the government for their product which uses the ML component as a kind product feature. In the sense of meta-responsibility, the state has the task of governance by establishing control frameworks and thereby establishing regulations such as those are already in place, for the protection of privacy (GDPR, 2016). The same applies, on supranational level for the European Union.

But can transparency and accountability also ensure that decisions do justice to ethical and moral considerations? Computers or AI models and algorithms do not have a *per se* built-in value system. According to Arendt (2007), models or algorithms are rather conscientious instances in the sense of obedient execution organs. According to Arendt (2007), morality arises only through a 'dialogue' with oneself and is also always related to oneself and the preservation of dignity. Hence it is about a person being a person not performing certain things, since she would no longer be at clean with herself. All this is completely alien to an algorithm. Therefore, only the people who design and use them can be made responsible for the morality of AI algorithms and models (Bostrom & Yudkowsky, 2014; Mittelstadt et al., 2016; Nissenbaum, 1996; Reichmann, 2019).

At the process level, however, essential questions remain unanswered and, for example, AI applications in domains like medicine, justice, personnel selection etc. cannot be done without the 'human entity in the loop' doing the final decision.

On European Union level a group of High-Level Experts was formed to define reasonable regulations and informal standards. This is also a task in various other countries and organizations (Algorithm Watch, 2020; Dutton, 2018; Library of Congress, 2019).

4.1. Ethical guidelines for trustworthy AI

The ethical guidelines for trustworthy AI mentions three main components, which should be met throughout the system's (using AI) entire life cycle:

- it should be lawful, complying with all applicable laws and regulations;
- it should be ethical, ensuring adherence to ethical principles and values;
- it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

In addition to this to realize trustworthiness AI should meet the following seven requirements, according to the guidelines:

1. Human agency and oversight: AI systems should enable equitable societies by supporting human agency and fundamental rights, and not decrease, limit or misguide human autonomy.
2. Robustness and safety: trustworthy AI needs algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of AI systems.
3. Privacy and data governance: citizens should have full control over their own data, while data concerning them will not be used to harm or discriminate against them.
4. Transparency: the traceability of AI systems should be ensured.
5. Diversity, non-discrimination and fairness: AI systems should consider the whole range of human abilities, skills and requirements, and ensure accessibility.
6. Societal and environmental well-being: AI systems should be used to enhance positive social change and enhance sustainability and ecological responsibility.
7. Accountability: mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. The framework for trustworthy AI is shown in scheme 1. All key requirements are interrelated between each other like shown in scheme 2.

Such guidelines for ethics of AI are surely not only an important task for European Union — many countries, especially the high industrialized nations started already with working on similar guidelines and rules for AI. The actual list of countries is shown in table 1.

4.2. Ethical guidelines and legal frameworks for trustworthy AI in the world

Table 1 shows which countries establishes an AI strategy in which year and if the paper or report also includes ethical guidelines for trustworthy AI. As one of the most comprehensive framework for ethical guidelines in the world the EU's AI Strategy can be seen. The leading economic nations like China, USA, India etc. all developed already an AI strategy with formulating regulations or a framework for trustworthy AI. Singapore e.g. describes a Model AI Governance Framework, Canada references on ethical norms. Japan describes principles of ethics in its AI Strategy. UK build up an own department the Centre for Data Ethics and Innovation within the Department for Digital, Culture, Media & Sport (Executive Office of the President, 2019, Future of Life Institute, 2020; Library of Congress, 2019).

4.3. Ethical guidelines and legal frameworks for trustworthy AI in Poland

Poland's government held its first roundtable on the development of a Polish AI strategy in May 2018. Attended by the Vice-President of the Council of Ministers, the Minister of Science and Higher Education Jarosław Gowin, the Deputy Minister of Digital Affairs Karol Okoński, and representatives of the scientific community and related institutions, the roundtable focused on the policies and tools needed to foster an environment conducive to the creation of AI technologies in Poland (Dutton, 2018). The Polish government will set up a range of observatories and specific chairs to tackle any ethical and legal issues Polish government to secure the creation of a trustworthy and sustainable environment for the development of AI. The idea is to monitor the international regulations and to make them usable for Poland and to raise recommendations for the legislative for reformation and set up new ethical guidelines. There is also the intention to support the recognition of interoperability standards and certification or compliance procedures of trustworthy AI (European Commission, 2020b).

4.4. Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is a new research scope on how to gain explainability and transparency within AI on an organizational or enterprise level resp. product level (Barredo Arrieta et al., 2020; Bejger & Elster, 2019). Using statistical methods and statistical learning, the current successful models developed are black box models. A black box model is a system that does not reveal its internal mechanisms and therefore in machine learning, 'black box' describes models that cannot be understood by looking at their parameters (e.g. an artificial neural network). The opposite of a black box is sometimes referred to as glass box. Interpretable or Explainable Machine Learning refers to methods and models that make the behaviour and predictions of machine learning systems understandable to humans.

Let us consider typical everyday business situation of loan application. The customer starts the interaction of a credit loan application and provides the needed details for the bank. The front-office captures the needed details and reiterates the steps if needed. The application is then assessed by providing the information to an AI model which is calculating, based on the input details, the possibility of a loan failure. There can be different models used, classification or a predictive model or even a combination of models. If the risk for credit fail prediction is above a certain defined threshold, the loan application request is denied. When the customer is asking for 'why' the application request was declined neither the front-office nor the developer is able to tell why as the model used to do the risk calculation is a black box model. As mentioned above a possible way during the design and development on process level could be the adaption of the CRISP–DM model by enrichment with the requirements of ethical guidelines and regulations.

The CRISP–DM (Cross-industry Standard Process for Data Mining, scheme 3) is the most common used process model for data mining/ machine learning. The process steps are business understanding, data understanding, data preparation, modeling, evaluation of the model and deployment.

To fully implement the requirements of the *Ethics guidelines for trustworthy AI* we suggest the adaption and enrichment of the CRSIP–DM model with the requirements defined by the guidelines (scheme 4). Most of the requirements have to be part of the process steps.

In the European Union, organizations who develop and use AI as well as the user of AI are subject to the regulations like GDPR (2016), non-discrimination and protection of privacy as well as consumer protection and product safety. The consumer therefore expects the same protection and respect of regulations when a product or service uses AI (e.g. European Commission ,2020a).

The enriched Cross Industry Standard Process for Data Mining (CRISP–DM) showed in scheme 4 is a standard process model with the intention to plan, to organize and to also execute a data mining analysis project. It can also be used for Machine learning purposes in the data analysis phase. Its steps are:

- business understanding,
- data understanding,
- data preparation,
- modelling,
- evaluation,
- deployment.

Starting with business understanding it is necessary to define the business problem and the objectives of the project. In this phase we see human agency and oversight, societal and environmental wellbeing as well as diversity, non-discrimination and fairness as a need (European Commission, 2020a). A business problem or objective which violates these requirements should be avoided as there is a high risk of potential damage for a single person as well as a group whether it is material or immaterial damage.

In the data understanding process-stage the needed data is to be analysed and it is to be decided which data needs to be collected in the case that the relevant data is not available. Without access to relevant data it is impossible to develop AI — therefore it is a part of the European data strategy to provide better access to that data and follow responsible data management methods and meet the requirements of the FAIR principles (European Commission, 2020c). A data exploration and a verification of the data is necessary. Beside these tasks the privacy and data governance considered to be fulfilled as only such data can be used which meets demand of the GDPR (2016) regulations. All findings within the stages business understanding and data understanding have to be documented to be auditable later on.

The data preparation stage is consisting of cleaning the data, generate new attributes, in the meaning of feature engineering and integration of data. Also, in this stage the ethical requirements mentioned above and GDPR (2016) regulations have to be fulfilled and documented.

In the modelling stage we perform our statistical modelling (e.g. design and train the model) and analysis. This stage consists of steps like selecting the appropriate modelling technique to use, design the test, build and assess the model. After selecting the right model for the problem, it needs to be trained avoiding overfitting, when it is a machine learning model and in the building process the appropriate hyperparameters need to be set. In this stage it is necessary, that the training data reflects no bias or discrimination. This could be the case e.g. by using inappropriate training data by prediction of recidivism regards skin colour, race or gender. The model needs to be documented and provided for the next step. At least the model needs to be assessed regards technical and business requirements. The model assessment is an information about the developed model and in case of many models the selection of the most appropriate one and the fine tuning of the parameters.

In the evaluation phase of the CRISP–DM process the results have to be investigated and it has to be decided whether these are good enough to deploy the model for everyday usage. The results have to be evaluated in regards of the business objectives defined in the first step. It also has to be tested in practical application environment for usage. This phase needs a documentation of the results how much they reach the business goals defined in first step.

If there are a lot of models used the approved models are those which meet the business criteria at the best. The whole process needs now to be reviewed by using the documentation of each step and again, checking if every step is meeting the ethical guidelines. If there are flaws or maybe not appropriates results or ethical requirements accountability and transparency (High-Level Expert Group on Artificial Intelligence, 2019) are violated or not fulfilled, the specific process step needs to be redone or the whole process.

In the last phase of the CRISP–DM process the deployment of the model needs to be conducted. This comprehends of several steps in the planning of the deployment, monitoring and maintenance, documenting and review of the results. By deploying the model in to greater use it has to be ensured that the risk of possible errors is minimized and if the model is a self-learning model it needs to be secured that the model developer takes care of the goals a model will be try to reach by self-learning in the future use. In specific critical domains, like law, autonomous driving, selection of personnel, medicine or loan approval a human should be always able to override the decision done by AI, to avoid false decisions done by the AI. In this final stage all documentations of all steps should be available.

4.5. Interpretability as a key factor of trust

Precisely speaking black box model belongs to the class of mathematical mechanisms which abstract a certain fragment of the real world (real decision mechanism *N*, named supervisor), which generates information *Y* on the basis of input data *X*. This information can be decisional indications. This concept is illustrated in scheme 5.

This mechanism consists in learning (selection of approximants of *N* from the set of hypotheses *H*) on the basis of data, in accordance with the algorithms named as learning machine *MU*, for the purpose of approximation of mechanism *N*. The approximated version of the mechanism *N*, which is hypothesis *h*, generates decision indications $\hat{Y}$ on the basis of a defined loss function, in order to achieve the best quality of information predication from the new input data (Vapnik, 2000). In black box model *MU* and *h* do not reveal their inner mechanisms and it is difficult to explain the way of generating input data $\hat{Y}$.

In chapter 4 we studied a very general notion of trustworthy AI defined by standards and regulations. However, there is an important question how to transform that postulates into operational guidelines for decision makers. We consider, according to current state of research that one of the major method to achieve trust is developing and operationalise the interpretability (explainability) quantitative methods. The adequateness of interpretability arises from an incompleteness black box problem formalization (Doshi-Velez & Kim, 2017), which means that for certain problems or tasks we need not only to get the decision proposal (prediction) but also we have to explain how the model came to the prediction. This scope of work seems to be particularly important in the context of applications in the economic domain, where not only a decision indication but also the method of its deriving must be absolutely comprehensible and transparent not only for the organisation internally (the decision-maker), but first of all for all of the stakeholders (customers, shareholders, regulators). There is a of course a problem of 'proprietary' type black box model (Rudin, 2019) but in we will not discuss that matter here, assuming that even the most confidential proprietary model should be interpretable — ready (for the case of authorities' request, for example). Although the notion of interpretability is still being discussed in the given context (c.f. Lipton, 2018) and is often interchangeably used with the notion of explainability (i.e. Lundberg et al., 2019). If we consider operational applications of decision mechanisms, this term can be defined as a skill of explanation or presentation of a model and decision indication in a manner which is comprehensible to a human (Doshi-Velez & Kim, 2017). Interpretability of a model may be achieved by ensuring (c.f. Lipton, 2018) its:

- transparency, which mainly refers to *MU* and *h*;
- interpretability of output data *post factum* (post hoc explanation), which refers to $\hat{Y}$.

According to Barredo Arrieta et al. (2020) transparency consists in:

- simulatability,
- decomposability,
- algorithmic transparency.

As an example (c.f. Barredo Arrieta et al., 2020), the decision mechanism (model) of a single decision tree is entirely transparent because:

- it is simulatable: a human may independently carry out simulations and obtain prediction on the basis of a decision tree without engaging any mathematical supply base;
- it is decomposable: the model comprise straightforward decisional rules which do not cause any changes of data and maintain their clarity;
- it is algorithmically transparent: contains rules which are readable to a human, which explained the information learnt from the data and allow a direct comprehension of the predication process.

However, if a single tree model becomes an element of an ensemble (i.e. random forest), such an ensemble must be considered a black box model which does not possess the transparency features of a single learner.

It is necessary to point out that transparent models are considered to be directly (locally or globally) interpretable (c.f. Arya et al., 2019).

Interpretability of output data is understood as an explanation (of a prediction, for example) carried out *post factum* from the functioning of the model, and according to many authors (i.e. Arya et al., 2019; Barredo Arrieta et al., 2020; Lipton, 2018; Ribeiro et al., 2018) it can amount to application of one or a few methodologies from the set they belong to:

- verbal interpretation — means generating textual explanations which facilitate to explain the results from the model;
- visualisation of output data and/or transformation of input data — a lot of such methods in literature relate to the dimensionality reduction of the problem, and they enable visualisation which is interpretable by a human;
- local post-hoc interpretability — refers to a division of output data space and explanation of less complex subspaces of solutions, which are representative of the whole model. In many publications (i.e. Arya et al., 2019; Lundberg et al., 2019) the notion of post-hoc interpretability is directly connected with the possibility to explain an individual decision indication generated by the learning mechanism (e.g. explain of credit decision $\hat{y} \in \hat{Y}$ for a customer based on its record of input data $x \in X$ and hypothesis *h*). In our opinion, local interpretation also includes an explanatory method based on examples, which is sometimes treated as a separate method. It is based on a deliberate selection of examples of output and input data (single records $(x,\hat{y})$ which

are representative for the indications generated by a given model and enable better comprehension of the model itself, or on studying counterfactual examples;

- studying the feature importances of input data — it is based on indirect explanation of the inner functioning of the model though calculating the relevance (importance measure) of particular elements of the vector *x* (or transformation of these elements) for the output data. Due to their inner construction, some model mechanisms possess inbuilt measures of this type on a global or local level. For other mechanisms, certain 'algorithmic overlays' are developed. They enable to determine the feature importances. The study of the feature importances on a local level should be, in our opinion, treated as an immanent element of the local interpretability;
- the construction of a surrogate model — a surrogate model is usually a directly interpretable model which abstracts more complex model, e.g. LIME (Ribeiro et al., 2016).

It is also necessary to mention that the methods which are to ensure interpretability of the model are divided also with reference to their connection or the lack of connection with a concrete learning mechanism:

- methods which can be applied for various types of learning algorithms (model-agnostic methods), e.g. the above mentioned surrogate models, which enable to measure feature importances;
- methods which can be applied only for one type or class of algorithm (model-specific methods), e.g. VIM measure based on measuring entropy for algorithms based on decision trees.

Having considered the one of the goals of the study, which is examination of methods of constructing trust to economic decisions, there is plenty more or less precise directives formulated by the different entities how to achieve clear answer why a particular model generates a particular output. It seems, however, that due to specific characteristics of an economic decision problems such as:

- direct results for the final user who is facing informational and technological dominance of the decision maker (consumer decisions, e.g. credit scoring);
- short time devoted to decision making process (an extreme case is automated decisional mechanisms);
- aspects which are not discussed in this paper, but which are undertaken by e.g. Rudin (2019), and connected with encompassing inner mechanisms with trade secret (as opposed to decision mechanisms which support medical diagnostics, which cannot be justified with economic benefit in a competitive race);

that the most appropriate methods of constructing trust are quantitative methods of local post-hoc interpretability, which comprise mainly methods of measuring the importance of the elements of input data or explaining local mechanism of decision generation.

4.6 Selected methods of local interpretability

A set of available methods of local post-hoc interpretability (both specific and independent of the decisional mechanism) is and is growing fast. Methods which, after our research, are worth mentioning in the context of economic decisions are first of all the methods which describe the local behaviour of a model with the use of linear, weighted combination of input data (Ribeiro et al., 2018). These methods are also named as additive feature attribution method (Lundberg et al., 2019). It has been demonstrated that:

$$\sum_{i=1}^{M}\phi_i = f(x), \phi_0 = f(\varnothing), \tag{1}$$

where $x \in X$, $\hat{y} \in \hat{Y}$, whereas ϕ_i is a feature attribution assigned to the element x_i of vector x (a single feature) after removing this feature (or a group of features) from vector x.

Equation (1) assures that the sum of the feature attributions equals the original model's prediction.

Among the methods based on the additive feature attribution, it is necessary to mention:

- LIME (Ribeiro et al., 2016);
- DeepLIFT (Shrikumar et al., 2016);
- Relevance propagation (Bach et al., 2015);
- Maple (Plumb et al., 2018);
- Saabas (2019) method;
- QII (Datta et al., 2016);
- SHAP (Lundberg et al., 2019).

Apart from those methods we would like to point to other, potentially useful in the domain of economic decisions, methods of local post-hoc interpretability. They are:

- anchors (Ribeiro et al., 2018) — method based on extraction of rules 'if−>then' for local prediction;
- method based on so called counterfactual explanation — the theoretical explanation of the fundaments of this method are included in Wachter et al. (2017).

In our empirical evaluation we focused on Anchors method, as less known and tested than additive feature attribution methods.

4.7 Empirical use case

At first we shortly present the theoretical background of the Anchors method and then present empirical use case.

The main idea of this method is to explain the behaviour of models by means of precise decision rules such as 'if−>then', which are called anchors. Anchors are locally sufficient conditions to assess the value of prediction, with a high degree of certainty. The rule 'anchors' a prediction if perturbations of predictors within the explained value do not affect the prediction. It means that for instances on which the anchor holds, the prediction is (with high probability) always the same. Let us assume that we hypothesis h was selected in the process of statistical learning, which generates decision indications. Let us name it $f(X){:}X{-}{>}\hat{Y}$. Then, $f(x)=\hat{y}$ for $x \in X$ is the local value of prediction which must be interpreted as for its formation and stability.

Anchor (decision rule) A is defined as:

$$E_{D(Z|A)}\left[I_{f(x)=f(z)}\right] \geq \tau, A(x)=1, \tag{2}$$

where: x represents the instance being explained (e.g., one record in a data set), A is a set of predicates, i.e., the resulting rule or anchor, such that $A(x)=1$ when all feature predicates defined by A correspond to x's feature values, $f(x)$ denotes the model to be explained (e.g., an artificial neural network model). It can be queried to predict a label for x and its perturbations. $D(\cdot|A)$ indicates the distribution of neighbours of x, matching A, parameter $\tau \in [0;1]$ defines precision threshold. Only those rules which achieve the local precision as at least τ are considered as anchors.

It can be therefore stated that anchor is a locally universal rule (independent of perturbations within x), which enable to explain how prediction $f(x)$ is generated with a relatively high probability τ. The measures which stem from definition (2) and define the quality of anchor are:

- precision $Prec(A) \geq \tau$, defines as:

$$Prec(A) = E_{D(Z|A)}\left[I_{f(x)=f(z)}\right]. \tag{3}$$

Because of the fact that in order to find anchor with precision matching the definition (3) requires examining $1_{f(x)=f(z)}$ for all perturbations $z \in D(\cdot|A)$, which is impossible for continuous values of x or big sets X, a statistical definition of precision is used. This definition induces sampling from distribution $D(\cdot|A)$ until probability of achieving precision τ does not reach the threshold level $1-\delta$ for parameter $\delta \in [0;1]$, i.e.:

$$P\left(Prec(A) \geq \tau\right) \geq 1-\delta. \tag{4}$$

- coverage[1] $cov(A)$, defined as:

$$cov(A) = E_{D(Z)}\left[A(z)\right]. \tag{5}$$

Coverage (5) is probability with which anchor explains the samples from distribution of perturbations D. Because condition (4) can be maintained for a big number of potential anchors, the anchor preferred is the anchor (A) for which $cov(A)$ is maximized.

[1] The symbol $cov(A)$ may be misleading (indicate covariance). Still, we have attempted to maintain notation and the source by Ribeiro et al. (2018) in accordance.

Anchor method is independent of the class of mechanism *f*(*X*), however applies only for classification problems.

We applied Anchors method for local interpretation of decision in credit risk assessment. Specification of a decision problem is as follows. One of numerous financial services offered by an entity is credit services within consumer credits[2]. Applicants (individual customers) apply for a loan, and the entity (company) decides to offer it or not. The entity registers annually a certain number of 'wrong loans', i.e. those credits in which the borrower is overdue with payment or applies for changing the conditions due to his/her insolvency.

As business goal the management of the company wants to decrease the number of "wrong loans) in the entire volume of credits by improving the quality of credit decisions. Using historical data, the company wants to build a model which may be helpful in prediction whether the applicant will fulfil his/her liabilities, and which may help to understand what qualities of an applicant may be a threat to the loan. This model will be a direct decisional tool in the range of credit decisions for the applicant.

Implementing quantitative approach to a decision problem company wants to build a classification model of AI which will be optimised on the basis of a training data set and used to support decision making process in the area of approval or rejection of a loan application (decision indication can be implemented by the decision-maker or automatically). Company expects that, based on an adequate post-hoc method, it will be able to explain in an objective way why the loan application of a particular customer has been approved of or rejected. As an illustration of completing the goals in the example a set of data collected by entity Lending Club (2019)[3] has been used. The data set contains raw data about 42538 loan applications and the loan history since the end or resolution of the loan agreement. The set *Y* contained one feature (target variable) describing the current status of the agreement (paid, resolved, delayed payment, etc.). The set of input data *X* contained 44 features, out of which 9 were selected at the phase of data preparation[4], i.e.: *loan_amnt* (the loan value), *home_ownership* (the status of property ownership in use), *annual_inc* (annual revenue), *desc* (recoded description of problems with payment), *inq_last_6mths* (the number of requests for payment for the last 6 months), *revol_util* (percentage of arrays not paid on time in full), *last_fico_range_high* (credit rating 1), *last_fico_range_low* (credit rating 2), *pub_rec_bankrupcies* (number of consumer bancrupcies).

The target variable was recoded into two values: *bad_loan, good_loan. Bad_loan* category stands for 13% of all the values of target variable. The data set was divided into a training set — validation set, and testing set in ratio 75% and 25%. The implemented learning model *f*(*X*) was a random forest classifier.

[2] In this example terms such as consumer credit and loan will be used as synonyms as the legal basis of offering those is of no importance.

[3] Data set is freely avaiable at Kaggle (2019).

[4] Empirical examples were carried out in the environment of Python language, version 3.6.3

For a trained model *h* the precision of classification was achieved on the level: training set 0.912, test set 0.831. The decision mechanism was prepared for application. In order to execute target b), verification of local interpretation of post-hoc prediction was carried out. For this aim, Anchors method was used. A record of predictors was selected at random from a test set. This record (number 23) consisted of (in the above mentioned sequence): 16000.00; MORTGAGE; 83000.00; 400.00; 3.00; 73%; 543.00; 521.00; 0. The decision mechanism *f*(*X*) generates *bad_loan* decision indication for this input data and it means rejection of loan. Based on case 23, the following anchor *A* was assigned[5]:
Anchor: *last_fico_range_high*≤649.00 and *last_fico_range_low*≤645.00 and *loan_amnt*>15000.00 and *revol_util*>72.23, for which following value of measures was achieved for the entire data set:
Precision: 0.99, Coverage: 0.04 for the assumed δ=5%.

The example shows how anchors method can assure insights into a model's local prediction and its underlying reasoning. The result shows which features were taken into account by the model (*last_fico_range_high*, *last_fico_range_low*, *loan_amnt* and *revol_util* in this case). Any interested entity can use this rule to validate the model's behavior. What is more, a decision rule which is calculated is valid for 4% of the entire number of records of features' perturbations in the presence of the anchor. In those cases, the explanation is 99% accurate, meaning the displayed predicates are almost exclusively responsible for the predicted outcome.

If we take into consideration only the test set, the measures for the anchor are as follows:
Anchor test precision: 1.00, anchor test coverage: 0.04.

What is also interesting is the study of reduced anchors, the precision of which is slightly lower than the one assumed, but which have bigger coverage:
Partial anchor: *last_fico_range_high*≤649.00 and *last_fico_range_low*≤645.00,
Partial precision: 0.91, partial coverage: 0.49.

This rule clearly indicates that regardless of the presence of other independent variables, in almost half of the available records, it is the conjunction of adequate values of independent variables *last_fico_range_high*, *last_fico_range_low* that is the most significant in the mechanism which generates rejection decisions.

As our example shows, given the same perturbation space, the anchors approach constructs explanations whose coverage is adapted to the model's behaviour and clearly express their boundaries. Thus, they are faithful by design and state exactly for which instances they are valid. This property makes anchors particularly intuitive and easy to comprehend.

[5] In the range of implementation of Anchors method, a code from Ribeiro (2018) was used.

5. Conclusion

AI provides a huge amount of potential to support and automatize economic decisions and therefore make them more efficient. To utilize AI and its chances and promises it is essential to fulfil the request for transparency, understandability or explainability and accountability by taking actions. The models and algorithms of AI used in daily products or life situations need to respect ethical and legal requirements. In critical domains like medicine, law, selection of personnel, loan application etc. the final decision still needs a human entity in the loop and need a proper tools to achieve a trust. The most comprehensive way to assure a significant level of trust is to transform law regulations (government, corporate, industry wide) into working schema of business process which could be applied universally in a decision making practice. We proposed a modification of standard CRISP DM process as an example of realisation of this approach. There is a need to operationalize such a process in a sense of transformation of directives and goals into methods to achieve them. As one of the operational elements in such general process we suggest implementation a viable local interpretability method. We underlined the connection between ensuring the trust (in very general sense) to economic decision generated by AI and practical, quantitative tools of local interpretability which can be applied. We examined one of them, the Anchors, and pointed to others which could be useful in economic decision making. We can state that Anchors has a very attractive features, making it a potentially straightforward for applications and use. The following are worth noting:

- easy to understand output in a form of rules which are interpretable by a human;
- the coverage measure plays a role of an importance measure;
- the method is model-agnostic and thus applicable to any classification model;
- the method is suitable for outputs that are highly complex in a neighbourhood of explained instance.

To summarise, only by fulfilling all requirements on the different stakeholder levels and over the complete lifecycle of an AI model a sustainable benefit by using AI can be utilized. There is a need to deliver regulation — consistent, operational procedures to ensure a trust for AI decision and our results could be taken into consideration in that scope.

References

Algorithm Watch. (2020). *AI ethics guidelines global inventory*. Retrieved 29.03.2020 from https://algorithmwatch.org.

Anderson, M., & Anderson, S.L. (2007). Machine ethics: creating an ethical intelligent agent. *AI Magazine*, 28(4). doi:10.1609/aimag.v28i4.2065.

Arendt, H. (2007). Über das Böse: Eine Vorlesung zu Fragen der Ethik. *München-Zürich: Piper.*

Arya, V., Bellamy, R.K.E., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., & Zhang, Y. (2019). *One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques*. Retrieved 29.03.2020 from https://arxiv.org.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *Plos One*, 10(7). doi:10.1371/journal.pone.0130140.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58. doi:10.1016/j.inffus.2019.12.012.

Bejger, S., & Elster, S. (2019). Das blackbox problem: Künstlicher Intelligenz vertrauen. *AI–Spektrum*, 1.

Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish, & W.M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139046855.020.

Brynjolfsson, E., & McAfee, A. (2011). *Race against the machine: how the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Lexington: Digital Frontier Press.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.R., & Wirth, R. (2000). *CRISP–DM 1.0: step-by-step data mining guide*. Retrieved 29.03.2020 from https://www.the-modeling-agency.com.

DARPA. (2016). *Explainable artificial intelligence (XAI)*. Retrieved 27.03.2020 from https://www.darpa.mil.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In *Proceedings of the 2016 IEEE symposium on security and privacy*. San Jose: IEEE. doi:10.1109/SP.2016.42.

de Laat, P.B. (2018). Algorithmic decision-making based on machine learning from big data: can transparency restore accountability. *Philosophy & Technology*, 31(4). doi:10.1007/s13347-017-0293-z.

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. Retrieved 29.03.2020 from https://arxiv.org.

Dutton, T. (2018). *An overview of national AI strategies*. Retrieved 27.03.2020 from https://medium.com.

European Commission. (2020a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European strategy for data* (COM/2020).

European Commission. (2020b). *National strategies on artificial intelligence: a European perspective in 2019: country report: Poland*. Retrieved 12.05.2020 from https://ec.europa.eu.

European Commission. (2020c). *White paper on artificial intelligence: a European approach to excellence and trust*. Retrieved 27.03.2020 from https://ec.europa.eu.

Executive Office of the President. (2019). *Maintaining American leadership in artificial intelligence* (E.O. 13859). Retrieved 29.03.2020 from https://www.federalregister.gov.

Ford, M. (2015). *Rise of the robots: technology and the threat of a jobless future*. New York: Basic Books.

Future of Life Institute. (2020). *National and international AI strategies*. Retrieved 12.05.2020 from https://futureoflife.org.

Gunkel, D.J. (2018). *Robot rights*. Cambridge–London: MIT Press.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy*. Retrieved 27.03.2020 from https://ec.europa.eu.

Holzinger, A. (2018). Explainable AI (ex-AI). *Informatik–Spektrum*, 41(2). doi:10.1007/s00287-018-1102-5.

Kaggle. (2019). *Loan statistics Lending Club*. Retrieved 21.10.2019 from https://www.kaggle.com.

Kurzweil, R. (2005). *The singularity is near: when humans transcend biology*. New York: Penguin Books.

Lending Club. (2019). *Loan statistics*. Retrieved 19.10.2019 from https://www.lendingclub.com.

Library of Congress. (2019). *Regulation of artificial intelligence in selected jurisdictions*. Retrieved 29.03.2020 from https://www.loc.gov.

Lipton, Z.C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10). doi:10.1145/3233231.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.I. (2019). *Explainable AI for trees: from local explanations to global understanding*. Retrieved 29.03.2020 from https://arxiv.org.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society*, 3(2). doi:10.1177/2053951716679679.

Molnar, C. (2020). *Interpretable machine learning: a guide for making black box models explainable*. Retrieved 27.03.2020 from https://christophm.github.io.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2(1). doi:10.1007/BF02639315.

Plumb, G., Molitor, D., & Talwalkar, A.S. (2018). Model agnostic supervised local explanations. *Advances in Neural Information Processing Systems*, 31.

Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR) (OJ L 119).

Reichmann, W. (2019). Die Banalität des Algorithmus. In M. Rath, F. Krotz, & M. Karmasin (Eds.), *Maschinenethik*. Wiesbaden: Springer. doi:10.1007/978-3-658-21083-0_9.

Ribeiro, M.T. (2018). *Anchor experiments*. Retrieved 12.05.2020 from https://github.com.

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Why should I trust you: explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. New York: ACM.

Ribeiro, M.T., Singh, S., Guestrin, C. (2018). Anchors: high-precision model-agnostic explanations. In *Proceedings of the thirty-second AAAI conference on artificial intelligence*. New Orleans: AAAI.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5). doi:10.1038/s42256-019-0048-x.

Saabas, A. (2019). *Treeinterpreter Python package*. Retrieved 07.07.2019 from https://github.com.

Sauerwein, F. (2019). Automatisierung, Algorithmen, Accountability. In M. Rath, F. Krotz, & M. Karmasin (Eds.), *Maschinenethik*. Wiesbaden: Springer. doi:10.1007/978-3-658-21083-0_3.

Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). *Not just a black box: learning important features through propagating activation differences*. Retrieved 29.03.2020 from https://arxiv.org.

Turner, J. (2019). *Robot rules: regulating artificial intelligence*. Cham: Palgrave Macmillan. doi:10.1007/978-3-319-96235-1.

Vapnik, N.V. (2000). *The nature of statistical learning theory*. New York: Springer. doi:10.1007/978-1-4757-2440-0.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). *Counterfactual explanations without opening the black box: automated decisions and the GDPR*. Retrieved 29.03.2020 from https://arxiv.org.

Waltl, B., & Vogl, R. (2018). Increasing transparency in algorithmic decision-making with explainable AI. *Datenschutz und Datensicherheit*, 42(10). doi:10.1007/s11623-018-1011-4.

Wischmeyer, T. (2020). Artificial intelligence and transparency: opening the black box. In T. Wischmeyer, & T. Rademacher (Eds.), *Regulating artificial intelligence*. Cham: Springer. doi:10.1007/978-3-030-32361-5_4.

Yudkowsky, E. (2001). *Creating friendly AI 1.0: the analysis and design of benevolent goal architectures*. Retrieved 12.05.2020 from https://intelligence.org.

Acknowledgements

Author contributions: authors have given an approval to the final version of the article. Authors contributed to this work equally.

Funding: this research was funded by the Nicolaus Copernicus University in Torun, Faculty of Economic Sciences and Management statutory sources.

Note: the results of this study were presented at *38th Prof. Bukietyński National Conference Methods and Applications of Operational Research* (October 20–22, 2019, Ciechocinek, Poland).

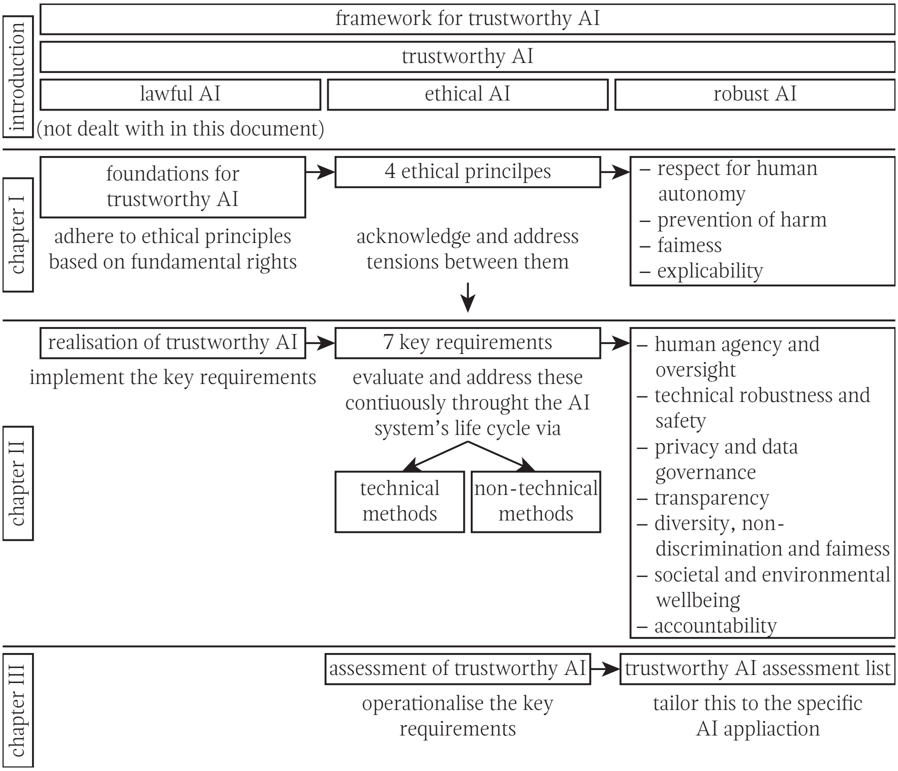
Appendix

Table 1.
Comparison of regulations of AI in different countries

Country	Report/paper	Year of release	Maturity of regulations	Has AI strategy	Role of trust and explainability methods in regulations
Canada	Pan-Canadian AI strategy	2017	high	yes	yes: ethical norms
Japan	AI technology strategy	2016	high	yes	yes: principles of ethics
Singapore	National AI strategy	2019	high	yes	AI ethics council/Model AI Governance Framework
China	New generation AI development plan	2017	very high	yes	yes: ethical norms
Finland	Finland's AI strategy	2017	high	yes	yes: Aurora AI
Denmark	Strategy for digital growth	2018	high	yes	yes
Russia	Decree of the President of the Russian Federation on the development of AI in the Russian Federation	2019	high	yes	yes
Italy	AI at the service of citizens	2018	high	yes	yes
France	France's AI strategy	2018	high	yes	yes
UK	AI sector deal	2018	high	yes	yes: Centre for Data Ethics and Innovation
USA	American AI initiative	2019	high	yes	yes
Australian	Digital economy strategy	2017	high	yes	yes: ethics framework
South Korea	AI information industry development strategy	2016	high	yes	yes
India	National strategy for artificial intelligence #AIforAll	2018	high	yes	yes
Germany	Germany's AI strategy	2018	very high	yes	yes
EU	EU's AI strategy	2018	very high	yes	yes: guidelines for trustworthy AI
Poland	AI development policy in Poland for 2019–2027	2018	high	yes	yes

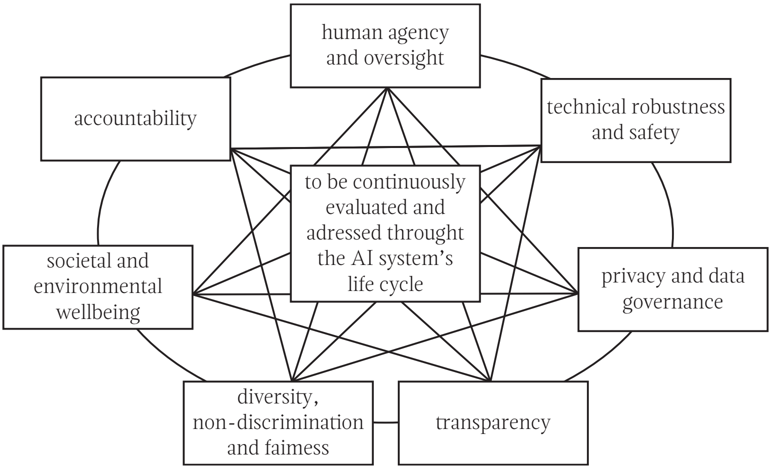
Source: Own preparation based on based on Algorithm Watch (2020), Dutton (2018), Future of Life Institute (2020).

Scheme 1.
The guidelines as a framework for trustworthy AI


introduction
framework for trustworthy AI
trustworthy AI
lawful AI
ethical AI
robust AI
(not dealt with in this document)
chapter I
foundations for trustworthy AI
4 ethical princilpes
– respect for human autonomy
– prevention of harm
– faimess
– explicability
adhere to ethical principles based on fundamental rights
acknowledge and address tensions between them
chapter II
realisation of trustworthy AI
7 key requirements
– human agency and oversight
– technical robustness and safety
– privacy and data governance
– transparency
– diversity, non-discrimination and faimess
– societal and environmental wellbeing
– accountability
implement the key requirements
evaluate and address these contiuously throught the AI system's life cycle via
technical methods
non-technical methods
chapter III
assessment of trustworthy AI
trustworthy AI assessment list
operationalise the key requirements
tailor this to the specific AI appliaction

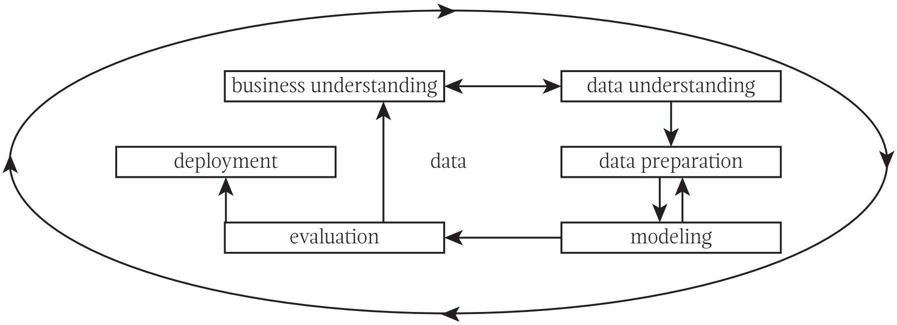
Source: High-Level Expert Group on Artificial Intelligence (2019).

Scheme 2.
Requirements for trustworthy AI


human agency and oversight
accountability
technical robustness and safety
to be continuously evaluated and adressed throught the AI system's life cycle
societal and environmental wellbeing
privacy and data governance
diversity, non-discrimination and faimess
transparency

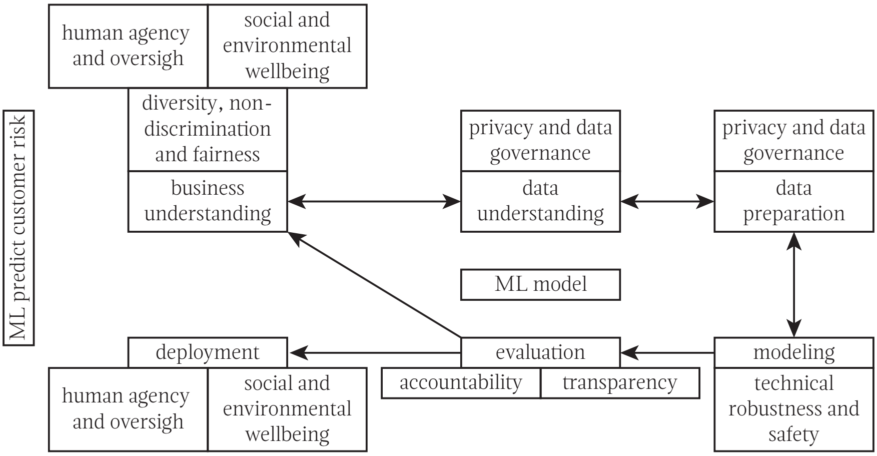
Source: High-Level Expert Group on Artificial Intelligence (2019).

Scheme 3.
CRISP–DM standard process



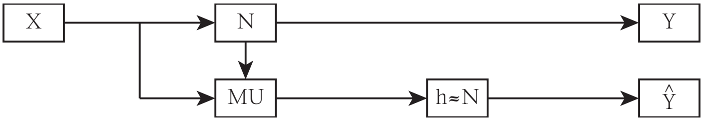
Source: Chapman et al. (2000).

Scheme 4.
Enriched CRISP–DM process model


human agency and oversigh
social and environmental wellbeing
diversity, non-discrimination and fairness
business understanding
privacy and data governance
data understanding
privacy and data governance
data preparation
ML predict customer risk
ML model
deployment
human agency and oversigh
social and environmental wellbeing
evaluation
accountability
transparency
modeling
technical robustness and safety

Source: Own preparation.

Scheme 5.
Machine learning by example



Source: Own preparation based on Vapnik (2000).