Chapter 2
Fundamental Fallacies in Definitions of Explainable AI: Explainable to Whom and Why?

D. O. Chergykalo and D. A. Klyushin

Abstract There are many articles that show a discrepancy between the various motives for the construction of XAI (Explanatory Artificial Intelligence), which is not surprising, since this area began to be actively centralized and actively developed only 6 years ago. But the strange thing is that the motives not only do not converge but may contradict each other. This indicates that there are fundamental errors in the very construction of different XAI concepts. These errors create not only contradictions between different visions of XAI, but also common to many concepts of error. The main one is the absence or incorrect answer to the question "For whom exactly should AI be explained?". Turning to human psychology and social processes that are accompanied by the exchange of explanations, we will try to consider what benefits the explanation brings to people and groups. Correcting fundamental errors in the construction of XAI concepts, we show that neural networks are no less explanatory AI than linear models and decision trees. Moreover, we will show what the neural network approach can do so that the explanation will not need to be exchanged for the quality of AI algorithms, and that they can even improve them.

Keywords Explainable artificial intelligence · Biomorphic artificial intelligence · Black-box models · Mental models

2.1 Introduction

2.1.1 A Short History of Explainable AI

The question of the explainability of AI did not arise due to the development of interest in AI bur rather the explanation of how the brain and psyche guided the development of AI. The earliest can be considered connectivism. Various psychologists have tried

D. O. Chergykalo · D. A. Klyushin (✉)
Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
e-mail: dmytroklyushin@knu.ua

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_2

to make psychology a more objective and constructive science, and therefore turned to neurobiology to link their vision of the psyche with the processes of the brain. An example is Sigmund Freud's "Project of Scientific Psychology", compiled in 1895 (Bob 2015). Obtaining information in this case is seen as a complication of the network, and learning is seen as the effective transfer of relationships in information to relationships in the network. And the vision of the brain as a multilevel distributed system was started in 1869 by neurologist John Hulingson Jackson (Greenblatt 1999).

These approaches were quite useful, based on them Friedrich Hayek built a model of learning Hebb's synapses (Hayek 1952) and hypothesized that the brain can be represented as a system of Hebb's networks (Hebb 1949). After that, this theory was developed and after some time, when computers became available to scientists, the first neural network model was programmed, that is the perceptron.

The very theory of networks as systems that can effectively adapt to external situations has proved its worth quite quickly. Pioneer of AI and cybernetics Joseph Licklider did not consider existing AI systems to be human-replacing, because as a psychologist he decided to improve the human–computer interface using mental models: ideas, strategies, ways of understanding that structure the existing experience (Licklider 1960). As early as 1962, in his work "Intergalactic Computer Network" (Licklider 1963), he described almost all the features of the modern Internet, and helped the Defense Advanced Research Projects Agency (DARPA) to create a prototype of the Internet: ARPANET. After that, improving the interaction between people and programs has finally become a long-term prospect for DARPA.

As Licklider identified, AIs at the time, that is, neural networks, could not replace some human functions. And when other researchers published articles explaining why this is so, investment in AI were frozen altogether. Researchers have tried to return to earlier ideas about the mind, which were started by Aristotle in ancient Greece. They began to build AI systems in such a way that they effectively used some rules of inference or symbolic manipulation, thus modeling some human dynamics within a single mental model. Although such systems were more understandable to end users, it quickly became clear that they were very fragile and could not evolve.

With the increase in the capacity of computers, neural networks began to develop again and gradually automate existing processes. But the expansion of the use of AI is gradually beginning to affect critical areas for people and humanity. At the same time, more advanced AI architectures are emerging, and it is becoming clear that rule-based approaches in mental models and neural networks do not contradict each other and can be effectively combined This was well noticed in DeepMind (Graves et al. 2014), and based on these views built their AI, which won other programs in various board games: AlphaGo, AplhaGo Zero, AlphaZero, etc.

DARPA, seeing changes in opportunities, in 2015 began to gather scientists to discuss and develope its future program "Explainable Artificial Intelligence". Using existing ideas about the effective interaction of man and computer and the consultation of scientists, they determined that the construction of AI requires a combination of three types of specialists: specialists in machine learning, specialists in human–computer interface and specialists in psychology of explanations.

Applications from scientists for the program began to be collected in 2016, from the same year the term "explainable AI" has become steadily popular. Of course, DARPA is not the only one who has thought about explanable AI, there are many articles reviewing various conferences on explainable AI (Adadi and Berrad 2018) as well as groups of scientists trying to act independently, the most influential of them is FAT ML (www.fatml.org). But DARPA can be singled out for their centralization in this topic providing a single key term for the problem, verifying existing statements for clarity, as well as developing a single system for evaluating explanations. Therefore, the analysis of their research is the best way to understand what problems are in the explainable AI.

Unfortunately, the structure of the DARPA's program "Explainable Artificial Intelligence" did not provide a way out of Licklider's ideas about human–machine symbiosis, and therefore did not take into account both the degree of AI autonomy and the need not only to transfer mental models but also to implement them effectively in AI. The report on their activities shows that they themselves understand that their initial idea of a "universal interface" between humans and AI was not very successful (Gunning et al. 2021). During and after their program, many articles analyze their results, but they do not go beyond DARPA's views, and most often repeat their mistakes.

Hereafter, we will discuss why this happens, and what contradictions in the perception of explainable AI hinder research.

2.1.2 Diversity of Motives for Creating Explainable AI

The paper (Lipton 2018) highlights the following motives for the creation of explainable artificial intelligence (XAI):

1. Trust, i.e. providing such an explanation of actions that will provide confidence in the algorithm.
2. Causality, i.e. obtaining causal relationships in the considerations of the algorithm.
3. Transferability, i.e. using of explanations of the algorithm for more correct application of the output of old models in new situations. For example, the model may deduce the probability of death from pneumonia, but interpreting this probability as a level of need for immediate treatment can be threatening and lead to even more deaths (Caruana et al. 2015).
4. Informativeness, i.e. obtaining useful information from the model.

There are also different motives for the use of XAI (Adadi, Amina):

1. Justification of the AI decision.
2. AI control.
3. AI improvement.
4. Obtaining from AI data that are useful for research in this area.

Some researchers (Doshi-Velez et al. 2017) consider these motives as part of some usage scenarios of explainable AI, each of which needs to be optimized separately, but, unfortunately, not everything is so simple.

In the following, the internal contradictions of each of these motives and their contradictions between each other will be presented.

2.1.3 Internal Inconsistency of Motives for Creating XAI

At first glance, the most truthful explanations are credible. But first of all, it is not possible to determine the correctness completely, especially for a non-expert. Information is perceived by a certain person, adjusted to the format of his ideas and "verified" with their help. Thus we have that for different people trust can be caused by different factors:

1. For non-experts: the use of their domestic intuition
2. For experts: the relevance of explanations to their experience and intuition
3. For scientists: the use to explain the facts of theories known to them.

Each of these cases requires explanations of different levels of complexity and different formats of information (for example, populism may inspire confidence in non-experts but may irritate and distrust academicians). If the goal of XAI is to increase trust, it will be able to make good explanations of not the best decisions, knowing what the person knows and how he "verifies" the information. Including using well-known human cognitive errors, stereotypes, etc. Other articles (Merry, M.) also point out that trust cannot be the main factor for explainable AI.

Causality. At first glance, if AI has found a pattern between two characteristics, it does not mean that he "believes" that there is a causal link between them. There may be a third factor influencing these, and AI tries to detect it using available information, but to guess about this factor will not work trying to dig into the internal structure of AI.

There are researchers who try to get explanations by extracting them from the local parts of the neural network (Guidotti et al. 2018), believing that in this way it is possible to isolate a part of the "neural network thinking process", which is incorrect due to the above. The problem with this approach is also emphasized by other researchers (Del Giudice 2021).

Tolerability. XAI is most useful for ordinary users and not for scientists (Gunning et al. 2021). The latter have methods of using old AI for new tasks and other methods of effective implementation of old models in new areas. They also know better how to use their models. And for practitioners implementing AI systems to know how to do it, scientists may need to create instructions for using this AI so that it can be adapted to needs. For example, the optimization goal might be to minimize the expected number of patient deaths and decide to change the old strategy to improve the target (expected number of patient deaths), knowing the number of staff required

for different levels of patient care and the previous distribution of their deaths, based on the old strategy of resource allocation.

Informativeness. At first glance, informativeness is the ability to get as much useful information from the model as possible. Informativeness can also be defined as the benefit of the information that a person receives. Although both of these definitions sound the same, they are not. According to research (Gunning et al. 2021), if the task is simple enough, then additional explanations will only annoy the person, and if the task is time-limited and/or already has a high cognitive load, then a large number of recommendations can even reduce the person's performance. In addition, the benefits of recommendations may change over time. That is, information should be useful for a particular person at a particular time in a particular situation.

As you can see, even within one motif we have different inconsistencies and contradictions. Of course, there are even more contradictions between the motives themselves.

2.1.4 The Contradiction Between the Motives for Creating Explainable AI

Trust is primarily a subjective feeling and it has nothing to do with understanding the work of the model or understanding its effectiveness. It has to do with how well something fits into a format that a person understands. And more precisely, it is related to our attitudes that work with data in formats that are clear to us and that tell us to trust or not to trust. Moreover, if the attitude of distrust and the attitude of trust is activated, then a person can both trust and distrust the subject. Some researchers even consider trust and distrust as independent and autonomous concepts (Van De Walle and Six 2014).

We can trust our loved ones in their judgments even more than experts. In addition, trust can be evoked by sympathy, positive associations, and how familiar you are with this object, using such factors, advertising can even lead to more trust without explaining how their mechanics or drugs work. There are many things that the advertising does not show as it is very difficult to fit into a format that is understood by the target audience. For example, drug advertising will not go into detail about the effectiveness and contraindications, which is the most useful information for the end user because its purpose is to interest society in their product and create some initial level of trust. In other words, trust is not related to causality (as, for example, the real reasons for the effectiveness of drugs can not be communicated to the end user), nor to tolerability (as in different areas of human life has its own settings), nor especially with information.

Causality can give trivial reasons that are already clear to man, and therefore can not help the person. In turn, identifying non-trivial causes is almost impossible, as the interdependencies in AI do not indicate a causal relationship. That is, limiting AI

so that it calculates everything due to the logic of causality can not only worsen the effectiveness of AI but also its informativeness.

Informativeness is usually understood as additional useful information, that is, one that is not obvious. But the model can learn to extract useful facts only from some area of application. Thus, with one-sided AI training to obtain information, tolerability is reduced. How to fix it we will talk in Sect. 2.3.

2.1.5 Paradigm Shift of Explainable Artificial Intelligence

We have considered many motives for the creation of AI and showed the contradictions in themselves and between them. These contradictions follow from the very procedure of constructing definitions for AI:

1. A human has his own needs to interact with AI
2. A human states the solution of these problems as an element of explicable AI
3. A human tries to generalize this problem, but due to limitations in its competences it does so only within its field of knowledge, while stating that this is XAI.
4. A human adjust other definitions of XAI for themselves, still highlighting from them only what makes sense within a fixed area of knowledge.

As already shown, different groups have different needs: trust, causality, tolerability, informativeness. They also have different formats in which it would be more convenient for them to interact with AI, as well as different types of information. The reason for the contradiction in the motives for the development of AI was stage 4: each considered how other needs relate to their needs without understanding the problems of other groups, and therefore not seeing obvious contradictions between these motives and the need for more correct generalization.

Some researchers say we just have to live with it, and the only thing we can do is try to structure all this variety of definitions of explainable AI in order to at least slightly reduce the formal and legal problems that arise when introducing AI in critical areas (Amann et al. 2022). But such an approach does not solve, but only slightly weakens the systematic problem. And in order to solve it and not come into conflict with the existing needs that are displayed in different definitions of explainable AI, it is necessary to understand the very motives for creating explainable AI.

There are good works that study the motivations for the use of AI in specific areas by interviewing employees of a particular company who specialize in the application of explainable AI in this area (Gerlings et al. 2021). But since they have not developed an understanding of the process of creating needs from the root cause, they often omit some groups of people who also play a key role in creating AI and also need explainability in accordance with their needs.

At the very beginning, the motive for the creation of XAI is the same as the motive for the creation of AI—to improve life through automation and optimization of existing processes. Since our interaction with AI has already become part of our

everyday processes (such as recommendations in Google, Facebook, etc.), it also fell under the need to be optimized, and XAI is the optimization of this process.

Considering XAI as an AI that optimizes the interaction of AI with a person or group of people, we have several possible directions for the development of XAI:

1. Learning an AI to interpret the internal processes of other AI.
2. Teaching an AI to get explanations based on the internal states of other AI.
3. Extending the old AI and training its multi-task abilities.
4. Learning a new multi-task AI that will display the necessary results and explanations.

It is known that multi-task contributes to more efficient internal representations, which can even improve the target output of AI. The fact that this method of learning is leading in the following models of XAI is indicated by some articles (Gunning et al. 2021).

2.2 Proposed AI Model

2.2.1 The Best Way to Optimize the Interaction Between Human and AI

To understand how to optimally adjust the interaction between AI and humans, it is necessary to look at how it is arranged and how human intelligences optimize interation with each other. By Piaget (2001) everybody from the beginning of the period of cognitive development named "preparation and organization of specific operations" (approximately 2 years) unconsciously or consciously receives a mental model of the world. Thanks to it, it can operate not only with external objects but also with concepts about them. From the age of 2, a child can listen to simple stories by ear and say simple sentences, perceive simple mental models and pass them on. In the future, a person learns to effectively express these elements of the internal symbolic model in a way that is understandable to others, that is to externalize them. Being in different social groups, a person gets different skills of externalization of this model, when talking: in the family, with educators in kindergarten, with peers, etc. The same goes for adults. For example, for groups of owners of certain resources (for example: water, land) for effective cooperation there are mental models such as ARDI (Actors, Resources, Dynamics, and Interactions) (Etienne et al. 2011). Theory of explanation is studied by various variants of externalization of internal symbolic models in the form of various mental models.

The idea of using the analogy of human communication and mutual learning to improve the explainability of AI is not new (Gallina et al. 2020). What these studies often don't include is that there are different explanatory needs at different stages of AI development and deployment, requiring vastly different mental models and explanatory skills.

In previous studies, we have written about how people's skills are organized like a tree, and how to transfer them to AI in general (Chergykalo and Klyushin 2021). In our case, we will focus as much as possible on the skills of explanations. This can be imagined as a branch of the skill tree with its branching into different skills. We will show in Sect. 2.3 how to optimally grow this branch. And before that, we need to show how explanation skills differ from forecasting skills.

2.2.2 Forecasts Are not Necessarily Useful Information

As we wrote in our previous research (Chergykalo and Klyushin 2021), people, when trying to think consciously, conduct a set of simulations of the future, through which we improve our function of choice. This principle is quite universal and can be used to improve the choice of AI. But this can only work when the AI knows the goal it is optimizing to make optimal decisions and has information about the external environment to predict its response to its actions.

In the case of systems where AI needs to cooperate with humans, to model events within AI, humans become part of the external environment that needs to be influenced so as to improve targets. If AI has the ability to influence the system, it, in the case of existing knowledge about the system, will build a set of possible scenarios for the future (the choice between which depends on its impact) and choose the most optimal of them. But when AI has no influence, or is modeled as if it does not have it, then AI can only predict one future—one that would be without its influence. AI cannot optimize anything if it has no effect on this system.

For example, a model can predict the probability of dying from pneumonia and indicate that the lowest mortality is from asthma. But this figure is maintained only by aggressive treatment. Therefore, if the prediction of the model affects physicians and their managers in such a way that they pay less attention to asthma and more to patients with symptoms that are predicted to be more associated with mortality, then mortality will only increase (Caruana et al. 2015). It is very interesting that when learning to predict the future, researchers do not teach the model to solve some problems, but teach it to model real processes.

For example, researchers may think that information about whether or not a person will be convicted in the future may help identify the most dangerous people to re-educate them. But when giving preliminary information for AI about a person, such researchers often begin to wonder why AI begins to use stereotypes about people to predict a person's future sentence. Picking up these results, publicists begin by accusing AI systems of racism or other prejudices. But, not surprisingly, their predictions of the existing AI systems show the problems of people who do not know how to fairly (even within the same legislation) to judge other people.

It is easy to cite examples of the fact that many characteristics of a person that he has throughout life and that are not related to his behavior subconsciously influence the decisions of others, including the decisions of judges and juries. The simplest such characteristic is attractiveness. There is good literature showing that people with

good looks has more chances for leniency (Castellow et al. 1990; Downs and Lyons 1990). Some studies (Stewart 1980; Beaver et al. 2019) show that attractiveness can halve the likelihood of imprisonment. Also, attractive people on average pay twice less fines in court for negligent damages (Kulka and Kessler 1978). But these are just some of the multipliers that reduce the penalties for attractive people. Attractive persons have many more privileges, such as reducing the filing of lawsuits themselves, supporting attractive people during difficult times, etc. (Benson et al. 1976).

There are many psychological factors that only reinforce this arrangement, such as the crowd effect. If you know that the majority of people vote for a decision, then subconsciously you also want to vote for it. For example, in the case of severe punishment, it will be more likely to be imposed when jurors vote by show of hands than by secret ballot (Kerr and MacCoun 1985). Probably the most interesting thing is that in all the above cases, the jury will be fully confident that they have made independent and fair decisions and generate plausible explanations for their decisions.

Quite often, advocates of ethics want AI, even if it has some idea of the person, not to take into account some factors. This requirement is very interesting, as these people represent that people may also, if necessary, ignore some factors. Of course, this is not the case, because if a judge requires jurors not to take into account some information that puts the defendant in a negative light, the jury's decision will take this information into account even more (Lieberman and Arndt 2000).

It is clear that in order for AI to interact effectively with people for a common goal, it needs to understand how to properly influence other people's actions through the explanations and recommendations it makes for them. He needs to understand how the whole system is designed to anticipate the actions of different people at different "stages of its life" to help researchers better tweak it, to help programmers integrate AI systems into our daily processes, to help users better adapt and help to extract long-term useful information during their interaction. In the following, we will talk about this system and how the old measures of explained AI are related to the effectiveness of AI–human interaction and practical goals.

2.2.3 Criteria for Evaluating Explanations

Researchers have already developed several explanation scoring systems, the most studied being the explanation scoring system (ESS) (Gunning et al. 2021). It consists of three main blocks:

1. Functionality measures:
    1.1. Speed generation of explanation and its assimilation by a human)
    1.2. Type of modality (visual, textual, etc.)
    1.3. The content of the explanation (justification, examples, reasons, connections that influenced the result (effect relations)
    1.4. Request for additional explanations (natural language, multiple choice, drill-down, etc.)

2. Learning performance measures. Checking the model on a test sample
3. Explanatory effectiveness measures:

    3.1. Explanation satisfaction
    3.2. Explanation goodness
    3.3. Mental model understanding
    3.4. User-machine task performance
    3.5. Trust assessment.

But this model does not reflect a comprehensive assessment of how the processes of AI interaction with humans and groups of people are optimized. To obtain this assessment, we will divide this process into several stages:

1. Stage of system development
2. Stage of system implementation
3. Stage of individual adaptation of people under AI and AI under people
4. Stage of joint practical activities
5. Stage of obtaining new practical knowledge from AI solutions.

Next it is necessary to choose a criterion that fits the area in which AI interacts with humans. For example, we will take such an area as medicine. The criterion of its success is the acceleration of the reduction in patient mortality. This criterion dictates what we need as soon as possible and at a sufficient level of quality to pass the (1), (2), (3) stage and optimize the transitions between (4) and (5) stage. Thus, even such a subjective feeling as trust in AI becomes an important problem if we consider it as a barrier (at the stage of individual adaptation) to the introduction of effective AI methods in medicine. We can say that at this stage it is really desirable to take into account human psychology (explanation satisfaction, trust assessment) so that a person begins to see in AI his colleague. And then, when it is necessary to make practical decisions we must be adjusting to the level of additional cognitive load that a person can afford to focus on the effectiveness of the transfer of mental models. Also, by analyzing the results (how much better/worse the patient ended up), it is vecessary constantly adapt the interaction with fellow doctors for greater efficiency (discussing existing situations, taking recommendations from doctors and giving recommendations to themselves).

If doctors do not perform surgeries or surgeries and, for example, only consult patients, the algorithm can not only give their view of the situation but also try to identify errors in the doctor's explanations or point out potential errors. Of course, this requires a large dataset that should show which doctors with what level of experience make which mistakes. This is necessary so that the amendments made by XAI are not meaningless for this doctor. Surgeons, if they have a difficult choice of methods that they want to use during surgery, can consult with AI before surgery.

How to create the necessary dataset and choose the architecture of XAI, so as to teach him to give explanations and recommendations at each of the above stages, we will tell in Sect. 2.3. And before that, let's show why understanding to whom and why explanations are so important, and what mistakes previous researchers made.

2.2.4 Explainable to Whom and Why?

Recently, researchers (Ribeiro et al. 2016; Gunning et al. 2021) define and work with explanations as with a certain universal interface. But as has already been shown for different groups of people, XAI needs to give different explanations, this is necessary so that XAI can not only explain to doctors or other end users, but also to integrate new technologies more quickly and painlessly into vital areas. Therefore, the approach with a universal interface is not correct in the sense that the interface is not universal but only for the end user.

However, one interface is useful just be the fact that it open the way for further development. Of course, this is only the first step in XAI, where a researcher who studies how best to create an interface plays the role of a UI/UX designer in an area such as human–computer interaction (HCI). In this area, there are so-called user tests that designers use to determine how information is perceived by users, how they cluster it. Using this information, they determine the optimal interface that minimizes the cognitive load on the user. To do this, they display information so that it already has a typical clustering and shows information that does not exceed a certain limit of cognitive load.

Researchers of the explanation interface faced a problem that is not so pronounced in HCI-designers, even basic explanations from the black box are quite difficult to obtain. To do this, they used a number of non-experts to select the interpreted indicators and give them an explanation. But even at this level it is already possible to identify problems such as "Explainable to whom and why?" For example, there are articles on XAI (Ribeiro et al. 2016) that select visualization and explanation techniques to help AI professionals find feedback loops or data leaks by taxing this interface so that even a non-expert can navigate. This is certainly useful for programmers who are trying to integrate AI into existing systems in order to improve the performance of a particular AI model and know its strengths and weaknesses to better integrate. But this is of no use to researchers who are creating new AI architectures themselves.

The same paper (Ribeiro et al. 2016) gave an example of the erroneous inference of the CNN neural network and shows that the "attention" of the neural network at this time was not focused on the object. Similar information has been presented in this article in the form that it can be useful to both AI developers and physicians. But it will not help neither the developers nor physicians. It can only help in the second stage of AI implementation, for programmers who adjust the model to real processes. For AI researchers, i.e. those who select the optimal AI architecture, such information is of no use also. For example, using the Fast R-CNN architecture, which first defines the area and then determines the type of object, you can easily include attention of AI as a factor to control and optimize.

It is more obvious that unnecessary information will only strain the doctor and may confuse him. For example, the fact that the coefficients of attention are more in the background does not mean that the internal picture is completely ignored, some of its elements may still stand out. It will be difficult for the doctor to navigate in

such cases and what they should tell him. At best, the programmer who implemented the system can give their functions of assessing whether to trust the system or not, which may already be useful to the doctor.

Usually, this method is not optimal for the interaction of doctors with AI. First of all, some articles (Merry et al. 2021) point out the lack of literature on how to improve the team work of AI. Improving the exchange of mental models has already proven to be a practice that leads to increased efficiency of care (Page et al. 2016). Mental models transmitted in operating rooms are already well studied (Nakarada-Kordic et al. 2016), and therefore it is easy to understand in what format should be the explanation of XAI in these cases, which may reduce retraining for AI.

2.3 Proposed Architecture

2.3.1 Fitness Function for Explainable AI

Summing up Sects. 2.1 and 2.2 we see that:

1. The main purpose of XAI is to optimize the processes of human interaction and AI.
2. Explanations cannot be evaluated by themselves, it is always necessary to focus on a key criterion in each area of application (reduction of errors of judges, doctors and increase their effectiveness).
3. The task of optimizing the processes of human interaction and AI can be effectively divided into 5 subtasks related to the 5 stages of the life cycle of AI.
4. At each stage of the XAI should give different types of explanations for different groups of people and for different subtasks.
5. Some of the explanations can be given by experts at the previous stages of AI implementation, in this case the XAI will add its explanations “on top” of the basic ones.
6. Knowing whether experts will give explanations to experts at the next stages has a strong influence on the format of XAI explanations, which shows that the XAI implementation system should be immediately ready to be established.

Learning the XAI system to give explanations at each stage, and to different groups of people, causes a multitask with a large number of sub-tasks to explain each of which must be evaluated. Fitness function can be the sum of all errors from different subtasks with weights. Each weighting factor will be determined in advance by the extent to which the explanations from the experts in the previous stages satisfy the experts on this subtask. If the explanation from previous experts is completely satisfactory, you can not even teach XAI this subtask. The more XAI explanations on this subtask can be useful, especially for the target, the higher the coefficient.

2.3.2 *Deep Neural Network is Great for Explainable AI*

Often simple models are more trustworthy than others. For example, it is enough to deduce the coefficients of the linear model in the diagram and you can already have some idea of their work. But with so many variables, it's unclear how they relate to each other and how to single out abstract concepts that can help with interpretation. Some co-researchers additionally teach sparse linear models to obtain more perceptible learning outcomes. But there is a clear ineffectiveness of such approaches, which is also pointed out by other researchers (Ribeiro et al. 2016).

There are also many tools for visualizing decision trees, with the help of which it is possible to understand more abstract concepts. But although they are good at capturing monotonic properties of data, they do very poorly to non-monotonic abstract characteristics, which, in complex problems, arizes no less frequently than monotonic.

It is easy to understand that the way non-monotonic abstract characteristics are calculated is quite difficult to depict by any methods, and even more difficult to understand what they mean. The task arises to ensure that these characteristics are somehow interpreted in a way that is clear to us humans. And the most effective teaching methods that work best with information and most effectively display it in text form are neural networks.

Neural networks can be taught to interpret some of the characteristics of simpler AIs by explaining them. But the most linear models, decision trees, etc. do not know how to effectively extract abstract characteristics from the data. What we have is that if we have already agreed that the neural network is best taught to explain something, the best way to give it information is to use another neural network or give it initial data directly. Thus, we have a neural network that encodes information, and this information is processed by two other neural networks: one displays targeted decisions (classification, etc.) and the other displays comments to explain their decision and possibly some recommendations.

As has already been shown by other researchers (Ribeiro et al. 2016), that if there are coefficients of attention in the architecture of the neural network-decoder, then it is possible to know what plays what role no worse than with linear models. The fact that neural networks are less clear than linear models is a myth is also pointed out by other researchers (Lipton 2018) Talking about the so-called post-hoc explanations.

The fact that AI now needs to solve both the main task and the task of explaining their actions is multitasking. We will talk further about why this is even good for the effectiveness of the main task.

2.3.3 *The More Multitasking the Better*

In our previous research (Chergykalo and Klyushin 2021) we have shown that multitasking is part of the work of biological neural networks. In addition, we have shown

that it is useful. Even when the tasks are not related to each other, mutual learning can help to regulate the internal representations, which can even slightly increase the efficiency of each task separately (Romera-Paredes et al. 2012). As mentioned above, when tasks are related, neural networks can even better shape internal representations. And in order to form all the necessary set of ideas in the neural network, it is good to teach it to explain their actions for the maximum number of groups of people and for different cases.

Thus, it is desirable to teach the neural network during each of the five tasks and also slightly adjust the amount of text for explanations or simplify some parts of the mental models that it displays or, conversely, learn to focus on certain aspects of mental models. All this is necessary when a person is cognitively overloaded or, conversely, considers the task very simple and may not catch important details.

2.3.4 How to Collect Multitasking Datasets

The use of mental models due to the existing fixed format can reduce the need for the size of datasets. But the main problem is that the conclusion for different explanatory tasks must be consistent. Because it will be very strange and confusing to see a case where a model will tell one group one thing and another another. From the outside, it will even look like the XAI is lying and will activate people's distrust.

The simplest and most reliable way out of this situation is to form groups of experts from different fields, and when this group solves some tasks, it must give a comprehensive explanation of its decisions in different mental models. For example, how would they explain their decision to the ethics team, how would they explain their decision to a fellow doctor, how would they explain the possible problems of their decisions to AI specialists (who will then see these explanations from the AI itself), and so on.

It is best to have the best experts in these groups who will give the best solutions and the best explanations. XAI trained on such data will be of the greatest benefit. But most likely there will not be much data from them. Therefore, as another option, it is proposed to include experts with an intermediate level of qualification—the main thing is that the expert who is key in making the decision should be at a fairly high level. In this case, it may be useful to include in the training data on the extent to which a highly qualified team of experts has made their decisions.

2.3.5 Proposed Neural Network Architecture

Some of the five subtasks are divided into their subtasks. For example, at the implementation stage AI can be evaluated not only by programmers but also by ethics specialists. At this stage, the greatest focus is given to how AI works with data, so a good solution is to use the reciprocal part for both of these tasks. Other researchers

have also written that subtasks in multitasking should be broken down into an effective hierarchy (Zweig and Weinshall 2013). Below we demonstrate the hierarchy of skills of explanations which will serve as a basis for our tree-like architecture of a neural network (Fig. 2.1).

In order to better understand this architecture, we offer a simple example of what such a neural network might look like. Showing it in Fig. 2.2, we omit the details (sizes of convolution kernels, activation functions etc.).

For this neural network, we can take for example the following fitness function:

$$f = E^2 = {E_{pred}}^2 + \frac{1}{100}\left({E_{adv1}}^2 + {E_{adv2_1}}^2 + {E_{adv2_2}}^2 + {E_{adv3}}^2 + {E_{adv4}}^2 + {E_{adv5}}^2\right)$$

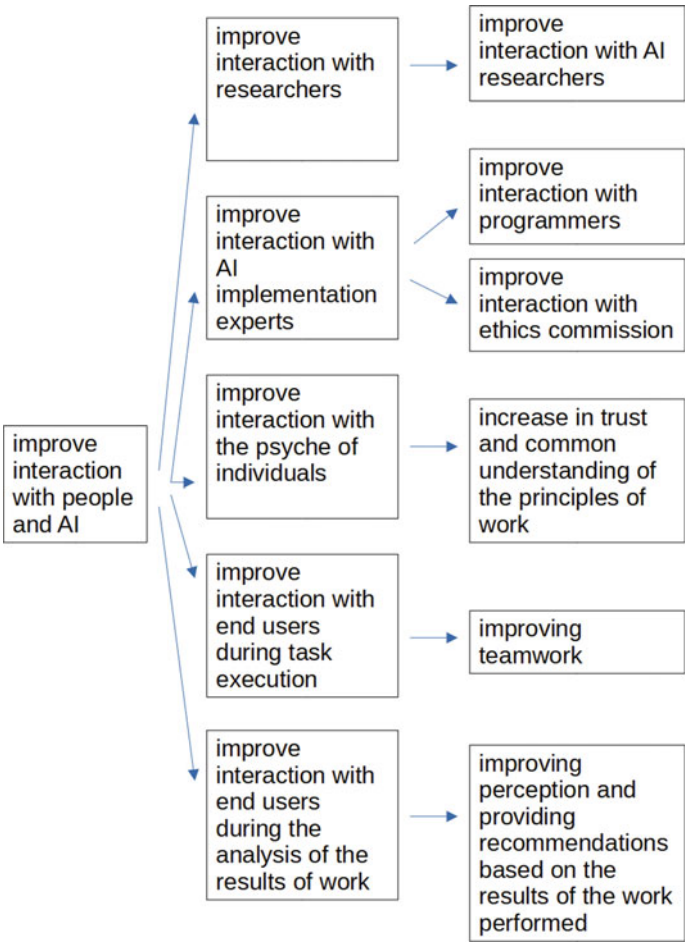

Fig. 2.1 Tree-like architecture of a neural network for XAI

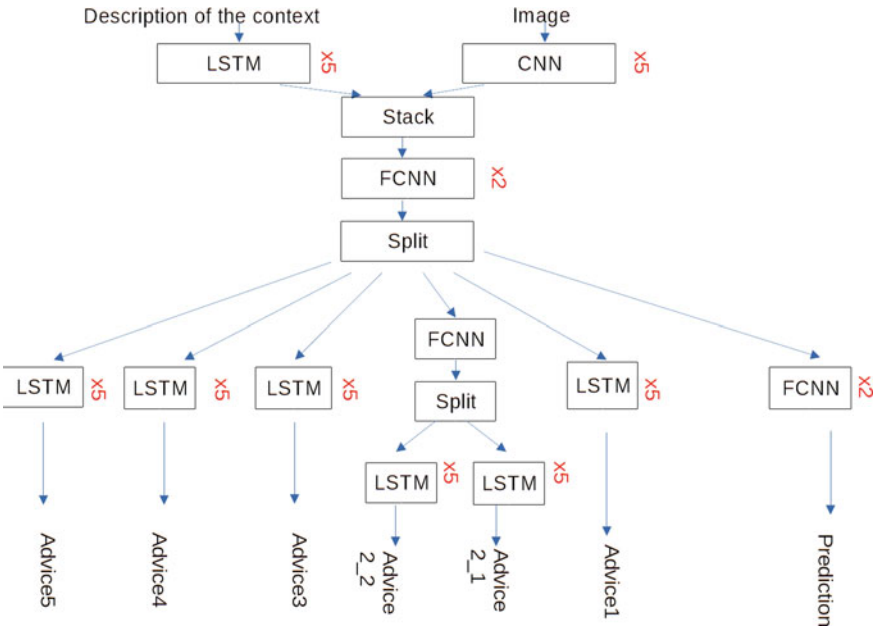

Fig. 2.2 A simple example of a neural network implementation. Xn—number of repetitions of layers

where—corresponding errors calculated for each neural network output.

We will also describe step by step how to adapt the one described in Fig. 2.1 architecture for solving domain problems on the example of medical image analysis:

1. Definition of versatile key data on the basis of which experts draw their conclusions. In the case of medicine, this is: a description of the context—a description of the mental state of the client and its background, a image—data obtained after the survey through medical devices.
2. Data collection as described in Sect. 2.3.4.
3. Choosing an appropriate architecture for encoding incoming data, transforming them and decoding the received pieces of information. For example encoding can be done for text data via Bert model and for images via ResNet convolutional layers.
4. Construct the fitness function formula in accordance with Sect. 2.3.1.
5. Train this neural network and implement its explanations in all areas described in Fig. 2.1. And thereby accelerate the creation, implementation, understanding and interaction with AI.

2.4 Conclusions

The effectiveness of XAI development is primarily determined by its implementation system. Therefore, there is a need for centralized and harmonized formats of mental models to be transmitted, as well as a centralized and harmonized XAI implementation policy. XAI optimizes the interaction between humans and AI, and the best and most experienced models of effective human-to-human interaction, and, by analogy, between human and AI, are mental models. We would also like to add that researchers avoid the phrase "learn from AI", but their needs and requirements, they show that this is one of the greatest values in these studies. Multitasking is the most promising area of XAI architecture. Taking into account this and other information, we have built a better and more comprehensive assessment of the effectiveness of XAI, which corrects errors in the purpose of XAI from past researchers. In addition, we have shown the general steps to implement a model that will satisfy this large set of requirements.

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Amann, J., et al.: To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digital Health **1**(2), e0000016 (2022)

Beaver, K.M., Boccio, C., Smith, S., Ferguson, C.J.: Physical attractiveness and criminal justice processing: results from a longitudinal sample of youth and young adults. Psych. Psychol. Law Interdiscip. J. Australian New Zealand Assoc. Psych. Psychol. Law **26**(4), 669–681 (2019)

Benson, P.L., Karabenic, S.A., Lerner, R.M.: Pretty pleases: The effects of physical attractiveness on race, sex, and receiving help. J. Exp. Soc. Psychol. **12**, 409–415 (1976)

Bob, P.: The brain and conscious unity: Freud's omega. Springer Science + Business Media (2015)

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, M.: Intelligible models for HealthCare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). Association for Computing Machinery, New York, NY, USA, pp. 1721–1730 (2015)

Castellow, W.A., Wuensch, K.L., Moore, C.H.: Effects of physical attractiveness of the plaintiff and defendant in sexual harassment judgments. J. Soc. Behav. Pers. **5**, 547–562 (1990)

Chergykalo, D.O., Klyushin, D.A.: Biomorphic artificial intelligence: achievements and challenges. In: Hassanien A.E., Taha M.H.N., Khalifa N.E.M. (eds.) Enabling AI Applications in Data Science. Studies in Computational Intelligence (Springer, Cham), vol. 911, pp. 537–556 (2021)

Del Giudice, M.: The Prediction-Explanation Fallacy: A Pervasive Problem in Scientific Applications of Machine Learning. PsyArXiv. December 13 (2021). https://doi.org/10.31234/osf.io/4vq8f

Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

Downs, A.C., Lyons, P.M.: Natural observations of the links between attractiveness and initial legal judgments. Pers. Soc. Psychol. Bull. **17**, 541–547 (1990)

Etienne, M., Du Toit, D.R., Pollard, S.: ARDI: a co-construction method for participatory modeling in natural resources management. Ecol. Soc. **16**(1), 44 (2011). https://www.ecologyandsociety.org/vol16/iss1/art44/. Accessed February 6, 2022

Gallina, B. et al.: Towards explainable, compliant and adaptive human-automation interaction. In: 3rd EXplainable AI in Law Workshop (XAILA 2020) co-located with 33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020) (2020). http://ceur-ws.org/Vol-2891/

Gerlings, J., Jensen, M.S., Shollo, A.: Explainable AI, but explainable to whom? arXiv preprint arXiv:2106.05568 (2021)

Graves, A., Wayne, G., Danihelka, I.: Neural Turing Machines. ArXiv, arXiv:1410.5401 (2014)

Greenblatt, S.H.: (1999) John Hughlings Jackson and the conceptual foundations of the neurosciences. Physis Riv. Int. Stor. Sci. **36**(2), 367–386 (1999)

Gunning, D., Vorm, E., Wang, J.Y., Turek, M.: DARPA's explainable AI (XAI) program: a retrospective. Appl. AI Lett. 2: e61. (2021). https://onlinelibrary.wiley.com/doi/full/https://doi.org/10.1002/ail2.61. Accessed February 6, 2022

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820 (2018)

Hayek, F.A.: The Sensory Order: An Inquiry into the Foundations of Theoretical Psychology. University of Chicago Press (1952)

Hebb, D.O.: The Organization of Behavior. Wiley, New York (1949)

Kerr, N.L., MacCoun, R.J.: The effects of jury size and polling method on the process and product of jury deliberation. J. Pers. Soc. Psychol. **48**, 349–363 (1985)

Kulka, R.A., Kessler, J.R.: Is justice really blind? The effect of litigant physical attractiveness on judicial judgment. J. Appl. Soc. Psychol. **4**, 336–381 (1978)

Lieberman, J.D., Arndt, J.: Understanding the limits of limiting instructions. Psychol. Public Policy Law **6**, 677–711 (2000)

Licklider, J.C.R.: Man-computer symbiosis. IRE Trans. Human Factors Electron. HFE-1:4–11 (1960)

Licklider, J.C.R.: Memorandum for members and affiliates of the intergalactic computer network. Adv. Res. Projects Agency (1963)

Lipton, Z.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (2018)

Merry, M., Riddle, P., Warren, J.: A mental models approach for defining explainable artificial intelligence. BMC Med. Inf. Decision Making **21**, 344 (2021)

Nakarada-Kordic, I., Weller, J.M., Webster, C.S., Cumin, D., Frampton, C., Boyd, M., Merry, A.F.: Assessing the similarity of mental models of operating room team members and implications for patient safety: a prospective, replicated study. BMC Med. Educ. **16**(1), 229 (2016)

Page, J.S., Lederman, L., Kelly, J., Barry, M.M., James, T.A.: Teams and teamwork in cancer care delivery: shared mental models to improve planning for discharge and coordination of follow-up care. J. Oncol. Pract. **12**(11), 1053–1058 (2016)

Piaget, J.: The Psycology of Intelligence. Routledge, London (2001)

Ribeiro, M., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (San Diego, California), pp. 97–101 (2016)

Romera-Paredes, B., Argyriou, A., Bianchi-Berthouze, N., Pontil, M.: Exploiting unrelated tasks in multi-task learning. Proc. Mach. Learn. Res. **22**, 951–959 (2012)

Stewart, J.E.: Defendant's attractiveness as a factor in the outcome of criminal trials: an observational study. J. Appl. Soc. Psychol. **10**, 348–361 (1980)

Van De Walle, S., Six, F.: Trust and distrust as distinct concepts: why studying distrust in institutions is important. J. Compar. Policy Anal. Res. Pract. **16**(2), 158–174 (2014)

Zweig, A., Weinshall, D.: Hierarchical regularization cascade for joint learning. Proc. Mach. Learn. Res. **28**(3), 37–45 (2013)