biomimetics

Article

Explainable AI: A Neurally-Inspired Decision Stack Framework

Muhammad Salar Khan [1,*], Mehdi Nayebpour [1], Meng-Hao Li [1], Hadi El-Amine [2], Naoru Koizumi [1] and James L. Olds [1]

[1] Schar School of Policy and Government, George Mason University, Arlington, VA 22201, USA
[2] Volgenau School of Engineering, George Mason University, Fairfax, VA 22030, USA
* Correspondence: mkhan63@gmu.edu

Abstract: European law now requires AI to be explainable in the context of adverse decisions affecting the European Union (EU) citizens. At the same time, we expect increasing instances of AI failure as it operates on imperfect data. This paper puts forward a neurally inspired theoretical framework called "decision stacks" that can provide a way forward in research to develop Explainable Artificial Intelligence (X-AI). By leveraging findings from the finest memory systems in biological brains, the decision stack framework operationalizes the definition of explainability. It then proposes a test that can potentially reveal how a given AI decision was made.

Keywords: explainable AI; interpretable AI; AI; decision stack; neurally inspired

Citation: Khan, M.S.; Nayebpour, M.; Li, M.-H.; El-Amine, H.; Koizumi, N.; Olds, J.L. Explainable AI: A Neurally-Inspired Decision Stack Framework. *Biomimetics* **2022**, *7*, 127. https://doi.org/10.3390/biomimetics7030127

Academic Editors: Isa Ebtehaj, Sayed M. Bateni, Babak Mohammadi and Stanislav N. Gorb

Received: 18 July 2022
Accepted: 7 September 2022
Published: 9 September 2022

1. Background

The recent crashes of two Boeing 737-Max commercial airliners have raised important questions about an embedded computational system (MCAS), which was installed to make the new 737 models feel more like the older models for human pilots [1]. Among the key issues raised is that the human pilots were not informed about the existence of the system and that the system's "intelligence" was subject to a single point of failure (an angle of attack sensor) [1]. Increasingly, Artificial Intelligence (AI) will play a significant role in such systems, particularly as autonomous machines operate in remote and hostile environments such as space or deep ocean [2]. In that harsh context, when failures occur, it will be essential to precisely assess what went wrong so that the designers can learn from failures. Meanwhile, when such systems make evidence-based decisions, explaining why and how a given decision was made is crucial. European Union law warrants such an explanation as part of the "Right to Explanation" enacted in 2016, mainly in the context of adverse decisions affecting citizens.

Modern AI systems operate on noisy and often uncertain data to make decisions on behalf of humans. When these systems work they are of great utility allowing for, among other things, self-driving cars and autonomous robots that operate in hostile environments. Beyond utility, these systems can also engage in self-teaching modes that allow them to excel beyond human capabilities at games like chess and Go [3–5].

However, as with human intelligence, sometimes AI fails to deliver. A well-known instance of such a failure is a Tesla Model S that was involved in a fatal crash while the car was in "self-driving mode" due to inaccurate feature extraction and intelligent comprehension of a white-colored truck by the AI [6]. The failure of AI is not surprising. Intelligence is the act of making decisions based on uncertainty. This fact differentiates AI from non-intelligent decision systems based on the flow-chart design, as in most computer programs [7]. For human beings, such failures are required for many kinds of learning during childhood and adulthood. Most machine learning (ML) AI algorithms also depend on a "training phase" whereby the artifact is instructed on a human-labeled dataset and learns from its failures before being allowed to operate in the "wild" on non-labeled data [8].

Therefore, it is understandable that, despite training, both humans and AI might mislabel a new instance of data that had never been seen or used before.

In the case of human intelligence, only recently has neuroscience offered a clearer picture of the cellular basis of learning and memory [9]. Furthermore, neuroscience provides evidence of a concrete hierarchy with the human body–brain system at the top and neuronal synapses at the base, allowing for a framework for explaining human decisions and their concomitant failures [10]. However, for AI, the explanation of why failures occur is not readily explainable [11]. This is in spite of European Union law requiring that such explanative AI be available to EU citizens to protect them from potential adverse effects of AI-based decisions such as the denial of credit [12,13].

Explainable AI (X-AI) is an artificial intelligence system capable of describing its purpose, rationale, and decision-making process in a manner that the average person can understand [14]. Here, we propose to advance the idea of a *decision stack* as a framework for explaining AI decisions, including failures. The term is beneficial as it reflects the idea that explanations must cross different levels of an organization in terms of complexity and abstraction [15]. In the next section, we will briefly lay out the literature that defines the field of X-AI and describe how our theoretical model fits in the field to explore a new way of looking at X-AI.

2. Explainable AI

The problem of explaining the outcome of a decision process is not new [16,17]. What makes X-AI fundamentally different is the added scale and dimensionality in modern AI decision systems compared to traditional decision trees or regression models [18]. This is where the problem of explainability becomes critical. As Paudyal puts it, X-AI is not an artificial intelligence system that can explain itself, but it is what we can try to interpret from the outcomes of an AI system based on our limited understanding of a process [19]. That is why some researchers use the word explainable AI interchangeably with interpretable AI. In other terms, how can we interpret an AI outcome to satisfy a specific question regarding the process that produced the outcome? It is important to note that in the attempt to understand the explainability of AI, we are not interested alone in the accuracy of the outcome of AI per se. In other words, not only do we want to explain why and how a wrong decision was made, but we also want to know why and how a correct decision was made.

Our interest in the process of decision arrival stems from the fact that a decision outcome could be true, but an AI system's process to reach that outcome might not be desirable based on our values. For example, an AI classifier that was designed to detect wolves among dogs, although it had perfect accuracy, based its decision on the presence of snow in the background of pictures with wolves [20]. In another instance, a law enforcement AI model designed to predict the risk of repeating criminal activity was found to have a strong racial bias [21]. AI systems can take shortcuts to arrive at a decision; however, these shortcuts are not always desirable. Therefore, the explainability of an AI system is separate from the accuracy of that decision system. This makes the attempt to develop an X-AI framework even more vital.

Various critical voices in the AI literature fundamentally challenge the concept of X-AI. We summarize these criticisms here in two groups. The first group believes AI is essentially too complex to be explainable. For example, researchers in the first group argue that the most popular AI systems have close to 100 million parameters, making it impossible to explain any specific outcome objectively; thus, this group believes that X-AI is a pointless endeavor [19].

The second group believes X-AI is possible but points out the challenge of the performance–explainability tradeoff. This tradeoff refers to the theory that in deep machine learning algorithms, more inputs and more hidden layers in a prediction model increase the accuracy of a model while interpreting the outputs becomes more challenging [19]. Thus,

they warn us that emphasizing too much on explainability might significantly damage the performance of a model.

In line with the first group's criticism, we acknowledge the complexity of AI. However, like most researchers, we believe that the question is not whether we can explain complex AI or not, but to what extent can we explain AI, what is a satisfactory explanation, and for what problem is the explanation needed [22]. As for criticism from the second group, we believe that line of criticism is concerned with the matter of design, i.e., how to implement a sufficient level of explainability to the design of an AI system. In our case, we focus on the problem of explaining an AI outcome post its design and development. Thus, the performance–explainability tradeoff is irrelevant here.

Explainability is crucial in policymaking and litigation [23]. Policymakers are especially expressing concerns over the emergence of "Blackbox" systems, which challenges the effectiveness of regulations [23]. That is why the "right to an explanation" movement, as mentioned, has been gaining momentum in policy circles, especially in Europe. Concurrently, the past few years have seen a plethora of literature on the importance of X-AI [24–27]. Parallel to the technical side of X-AI, a separate body of literature has been addressing explainability from the perspective of human rights, social justice, and fairness [28–31]. Felten has previously surveyed the X-AI literature and explains that any successful X-AI endeavor should help AI systems reach these four goals: transparency, accountability, safety, and fairness [32]. The X-AI framework presented in this paper aims to be used as a foundation for any domain-specific X-AI system to improve its transparency, accountability, safety, and fairness.

One final critical aspect of X-AI that has gained attention in the literature is the issue of evaluating the explainability power of an X-AI framework. Simply, how can we assess if an X-AI framework has adequately explained a decision process to an end-user? Scholars point out to the fact that the final goal of X-AI is eventually to convince a person about the credibility of an AI outcome [22]. This puts the end-user at the heart of X-AI. That is why 'the power to convince' is central to most X-AI evaluation frameworks. Hoffman et al. have proposed the most comprehensive guidelines for evaluating X-AI frameworks [25]. They built their guidelines based on the idea that "the property of 'being an explanation' is not a property of statements; it is an interaction" [25]. They also note that an explanation depends on end-users needs and their understanding of the AI outcome. Hoffman et al. essentially propose a qualitative effort using satisfaction surveys, questionnaires, mental models, and checklists.

While Hoffman et al. have successfully offered an evaluation model for X-AI frameworks, it relies heavily on end-users subjective understanding and satisfaction. Unfortunately, the literature lacks a universal objective framework that can explain an AI outcome independent of the end-user, i.e., an explanation that can be used in a court of law or to solve a policy problem where objectivity is desired. Our proposed theoretical framework in this paper aims to satisfy this need. However, before we offer our framework, we summarize existing X-AI methods and some of their issues below.

3. Current State-of-the-Art Explainable AI Methods and Approaches

Machine learning interpretable techniques (or X-AI methods) aim to understand ML models' decisions (predictions) and explain them in human terms to establish trust with stakeholders, including engineers, scientists, consumers (users), and policymakers. The field is not nascent [33]; its early days trace back to the origins of AI research in the development of human ML systems [34]. Since about 2015, there has been a resurgence in X-AI research that parallels the advance in increasing problems of applied ML systems in society. As a result, we have seen a suit of interpretable ML or X-AI methods particularly to untangle deep learning models [34]. One can choose from various ML interpretability techniques (shown in Figure 1) for any use case [34–37].

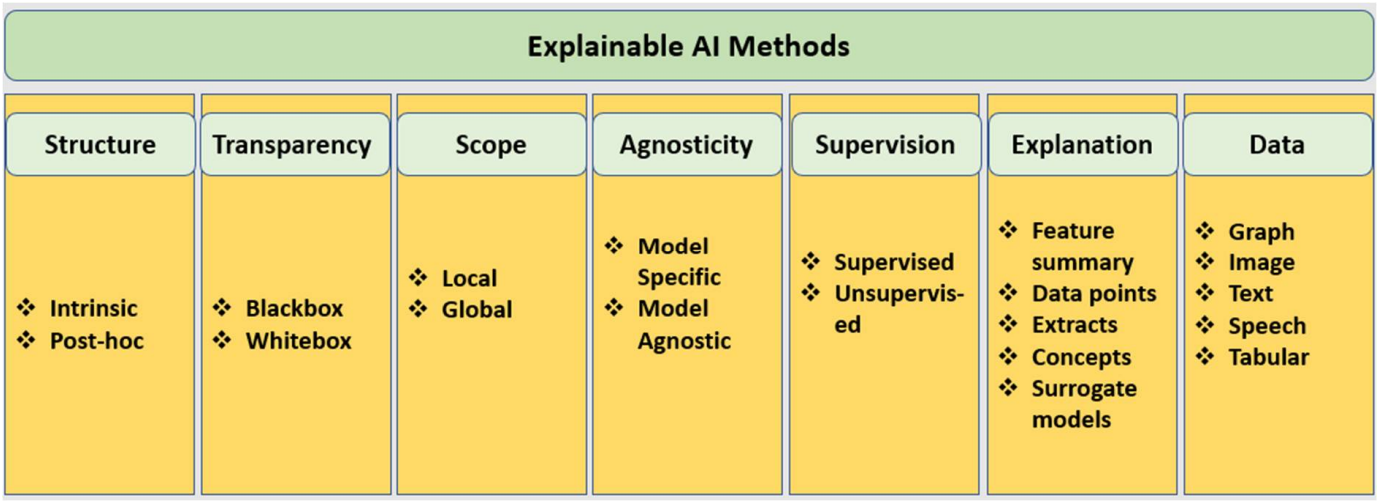

Figure 1. The figure lists various explainable AI methods.

Below, we summarize a landscape of the techniques (listed in Figure 1) from the literature based on certain criteria, including structure, design transparency, agnosticity, scope, supervision, explanation type, and data type.

3.1. Intrinsic vs. Post-Hoc: The Criterion of Structure

This criterion differentiates whether interpretability is achieved by containing the complexity of ML models, known as intrinsic (aka simple models), or by applying methods that analyze models after training (called post-hoc models) [35,38,39]. Intrinsic models are easily interpretable because of their simple structure [39]. Examples of the intrinsic model are linear regression, logistic regression, decision trees, and k-nearest neighbors. Post-hoc are complex structure models that achieve interpretability after model training [39]. Examples include permutation feature importance and neural networks [40]. Such models take into account changes in the feature or neural space and how these changes affect the outputs.

Generally, intrinsic models return simple interpretable explanations, but they lack in offering high-level predictions for complex problems [34]. On the other hand, post-hoc models perform better on most tasks but are too complex to understand for humans [34]. Neural network models, for instance, have millions of parameters that surpass human capabilities. These post-hoc models necessitate the need to derive human explanations for complex ML models.

3.2. Blackbox vs. Whitebox vs. Greybox Approaches: The Criterion of Transparency in Design

This criterion distinguishes based on what we know about the design of a method. A Whitebox approach is more transparent and explainable by design than a Blackbox approach [41]. Examples of Whitebox approaches include simple decision trees, rule-based models, patterns-based models, linear regression models, bayesian networks, and fuzzy cognitive maps [42,43]. Other methods following the Whitebox approach include fuzzy decision trees and fuzzy rules-based models. As opposed to Boolean logic (true/false statements, for instance), such algorithms follow a fuzzy logic (human-like reasoning with statements that could lie in a spectrum of truth/false) and hence take into account an uncertainty underlying analyzed data in explaining the decision [44–49].

As opposed to the aforementioned Whitebox methods, deep neural networks and random forests are some examples of Blackbox approaches [41,50]. Blackbox approaches usually contain complex mathematical functions like support-vector machine and neuronal networks; thus, they are generally hard to understand and explain [43]. On the other hand, most Whitebox models can be comprehended by experts as their models are closer to the human language [43].

In the Blackbox approach, we only know the relations between inputs and outputs and the response function to derive explanations [51]. As far as Whitebox approaches, we have access to the model-internal parameters [51]; in other words, we can access weights or gradients of a network. In general, Whitebox approaches are more interpretable but less accurate, whereas Blackbox approaches are more accurate but less interpretable [41].

Finally, there are Greybox approaches as well, which lie in between Blackbox and Whitebox models [41,52]. Examples of such approaches include Local Interpretable Model-agnostic Explanations (LIME) and Interpretable Mimic Learning [41]. In Greybox models, an expert knows when how some part of a system works mathematically (Whitebox) and is uncertain about the others (Blackbox). The Whitebox part of such models is fixed due to the underlying physical structure and constraints, whereas Blackbox part needs to be learned from the data. Greybox models, combining Blackbox and Whitebox features, may acquire the benefits of both, causing an explainable model which could be both accurate and interpretable simultaneously [41].

3.3. Local vs. Global: The Criterion of Scope

This criterion distinguishes methods based on whether the scope of the interpretability applies to the whole or part of the model. In local approaches, the scope of interpretability is limited to individual predictions or a small portion of the model prediction space [53,54]. On the other hand, global methods cover the entire model prediction space [53,54]. This is accomplished by aggregating input variables' ranked contributions towards prediction space (or decision space).

Local approaches provide a larger precision particularly of individual prediction (or a specific decision) but lower recall understanding of model behavior across all examples [54]. On the other hand, global approaches have a higher recall view of the model prediction (i.e., help one comprehend complete decision structure) but lower precision due to aggregations such as medians or means, which obscure individual contributions [54].

Some examples of local approaches are Local Interpretable Model-agnostic Explanations (LIME) [20]. SHapley Additive exPlanations (SHAP) [55], and Individual Conditional Expectation (ICE) [56]. Examples of global approaches include Partial Dependence Plot (PDP) and Accumulated Local Effects (ALE) [56,57].

3.4. Model Specific vs. Model Agnostic: The Criterion of Agnosticity

This criterion differentiates XAI methods on the level of agnosticity. Model agnostic methods refer that their X-AI algorithm can be applied to any kind of ML model [54,58]. Such methods do not depend on model internals. Instead, they rely on changes in input features or their values to understand how they influence the outputs of a use model. Examples include SHAP and LIME, which are portable across different model types [20,58]. Conversely, model-specific methods are designed for specific types of ML model [58]. Examples are methods that depend upon some intrinsic parts of model learning methods, such as neural network methods [54,58].

Agnosticity criterion could be a blurry boundary between various methods. One may aggregate the scores of some local model-specific methods (such as integrated gradients or SHAPLY) to revive the entire prediction space employing aggregation operations like averages and medians. Such methods would be termed as hybrid methods [59,60].

3.5. Supervision-Based Methods

This criterion distinguishes among methods based on the degree of supervision. Some examples of these methods are AI attribution methods, rationale-based methods, and disentanglement representations. While attribution methods entail an active manipulation of input data (supervised), rationale and disentanglement representations are unsupervised methods in the sense that researchers assume no explicit annotations about input data. Examples of attribution methods include LIME, SHAP, Integrated Gradients, SmoothGrad, Layer-wise Relevance Propagation, and Perturbation meth-

ods [61]. In rationale methods, pieces of input texts are extracted to determine a possible justification (Rationale) for prediction [62,63]. Researchers have no say in determining which words should be included in the Rationale [62]. Similarly, in disentanglement representations, latent-variable models learn representations of high-dimensional data in an unsupervised manner [64].

3.6. Explanation Type-Based Methods

There is a variety of interpretability methods that further differ in explanation output [35,58]. For instance, techniques such as feature summary return feature statistics, measuring a feature's proportional contribution to the prediction [65]. Similarly, other returning data points help us better understand the models [66]. Some different approaches help us build simple models around complex ones. Those simple models, called surrogate models, can be used to derive explanations [35]. Finally, other explanation type-based methods extract concepts, decision rules, correlation plots, and other visualizations [67–69].

3.7. Data Type-Based Methods

Besides all the above criteria, we can further differentiate according to the data type a method can handle [70]. Not all X-AI algorithms can work with all data types. Examples of data types may include graph, image, text/speech, and tabular [35,71–74].

The above paragraphs provide an overview of the various methods that have already been developed. Readers interested in more details can refer to other established and emerging literature [34–38,70]. Overall, each X-AI method has different guarantees, limitations, and computational requirements in explaining outputs. Nevertheless, these excellent foundational methods help create some model understanding and offer bits of human interpretable understanding. However, there is still so much for the scientists and engineers to understand how the AI implemented a decision while explaining the model decision to the public, policymakers, and regulators.

The methods are not a panacea to all our problems in the X-AI research field [75–77]. One issue discussed in the previous section is the interpretability–performance tradeoff [77–79]. For the last decade, as researchers and engineers have tried to increase performance or even exceed human-level performance through AI algorithms (for instance, via deep learning algorithms), the effort has costed reduction in the level of explainability.

High-profile ML deployment failures from these Blackbox models offer convincing evidence of the claim made in the preceding paragraph [80–84]. The failures point out that the models, particularly Blackbox (post-hoc) models and methods, are very opaque, uncontestable, exhibit unpredictable behavior, and in some cases reinforce undesirable racial, gender, and demographic biases [85]. All these affect crucial outcomes for the public, engineers, scientists, and policymakers. High-stake settings such as healthcare, criminal justice, and lending have already reported significant harms because of the problems inherent in Blackbox methods [85].

Another issue is the kind of explanation these X-AI methods return. So far, the explanation is incomplete (i.e., still model output is not fully predicted) [86,87]. Similarly, the explanation lacks accuracy. For example, four X-AI methods (including LIME) were deployed to predict a matchstick [86]. In other words, researchers were curious to know what makes an image of a matchstick a matchstick. Is it a matchstick because of the flame or a wooden stick? In their default settings, by changing a single parameter, the methods returned 12 unique explanations suggesting the methods are unstable in prediction [88]. The X-AI methods may also misdiagnose cancer if medical images can get modified in ways unknown to human understanding [89]. In a similar vein, the explanation must be meaningful and understandable. Unfortunately, X-AI methods have yet to make strides to return meaningful explanations. In one instance, deep neural networks mislabeled the image of a lion as a library [90]. All these issues might get resolved once the scientists thoroughly research and understand the intricacies within Blackbox models.

Finally, another vital challenge these methods face is the challenge of internal and external validation. Internal validation demands that we rule out alternative explanations, establish causality direction, and account for simultaneity and selection bias [91]. However, we have seen the internal validation of a method severely challenged by changing the input data (features or other data parameters). For instance, in the famous example of husky vs. wolf, the X-AI method (LIME as developed by Ribeiro et al. [18]) failed to predict a husky on the snow due to a bias in data. As opposed to the training set (a husky on the grass, a wolf in the snow), in the validation set the researchers flipped the usual background in some images (husky on the snow and wolf on the grass). Another instance of failed internal validity is when autonomous vehicles misread (mispredict) a slightly blurred stop sign [81], or, by changing a single pixel on an image makes an AI think a horse is a frog [82], or in the case of medical imaging, X-AI algorithm misclassify a brain tumor [89]. Similarly, AI methods failed to predict a school bus' right-side-up when the bus was rotated [80], suggesting AI's brittleness and weak internal validity. All these have repercussions for users.

Unlike the model's precision and internal mechanics, external validation would demand a successful application of X-AI methods from a small validation dataset to a larger population or on a range of data [91]. This is essential as AI moves from toy lab experiments and returns results on wild data in different circumstances. Existing X-AI methods suffer from external validity concerns and cannot handle all sorts of data. Some methods work with visuals; other techniques take only text and speech. The validity issue is even more problematic when the methods applied to the same data return different predictions, as elaborated in the example of matchstick prediction [86,92]. In the exact matchstick prediction, some researchers increased the sample size from 50 to 800, and the prediction space of heatmaps changed [93], suggesting the X-AI method is not robust when the sample size increases. Such external validity issues will be highly concerning as AI applies to human ML systems in healthcare, legal, defense, and security arenas.

There is an urgent need for an X-AI framework that explains the model and educates the users about AI decisions, building trust among various societal stakeholders. Such a framework will ideally return complete, accurate, consistent, reliable, and meaningful explanations comprehendible in various instances across multiple domains both in the lab and the practical world. In this concept paper, we look to the human brain for inspiration and derive a neurally inspired framework named as decision stack framework. Our goal for the framework is to employ any dataset, mimic human brain circuitry, and generate meaningful understanding, thus disclosing the AI decision Blackbox in its entirety. Furthermore, since the human brain is intelligent, chosen for throughout evolution, and universal [2,94], we believe the framework will be robust and efficient in prediction. In the following paragraphs, we elaborate on the lessons from neuroscience for motivation and explain the non-AI machine decision stacks for comparison before elucidating our neurally inspired framework.

4. Existence Proof: The Lessons from Neuroscience

Since the advent of non-invasive human functional brain imaging in the last decades of the 20th century, it has become commonplace to observe the brain correlates of conscious human subjects as they make decisions at a resolution approximately 1000-fold greater in space and time than the neural code itself [14,95–99]. These blurred images of human brains "caught in the act" of making intelligent decisions have been striking, albeit problematic, from the standpoint of reproducibility. Nevertheless, scientists have observed clear localized neural signatures of learning and memory [100,101]. Impressively, AI systems have been trained on such signatures and can correctly label them with appropriate nouns (e.g., cup, banana).

Blurred images of human brains in action have been connected to the lower levels of the human decision stack by numerous animal studies that recorded from individual nerve cells and even individual synapses [102,103]. These animal studies assume that

mechanisms of mnemonic function have been conserved across phylogeny. Recent work on human epilepsy patients supports this notion [102,103]. Operationally, neuroscience has coined the word "engram" to represent the cell assembly of neurons that participate in an individual memory [104]. These engrams can be visualized in animal models and correlated precisely to cognitive and behavioral states in the same experimental subject.

Most importantly, scientists have developed novel optogenetic techniques that enable experimenter-controlled switching on and off of specific engrams to produce corresponding amnesia and subsequent memory rescue in animal models, including mice [105]. Thus, the base of the decision stack, at the level of the cell assembly, has been connected to the top of the stack, at the level of the brain-body. The intermediate components of the neurobiological decision stack correspond to various cortical modules, brain nuclei, and their associated connections, as illustrated in Figure 2.

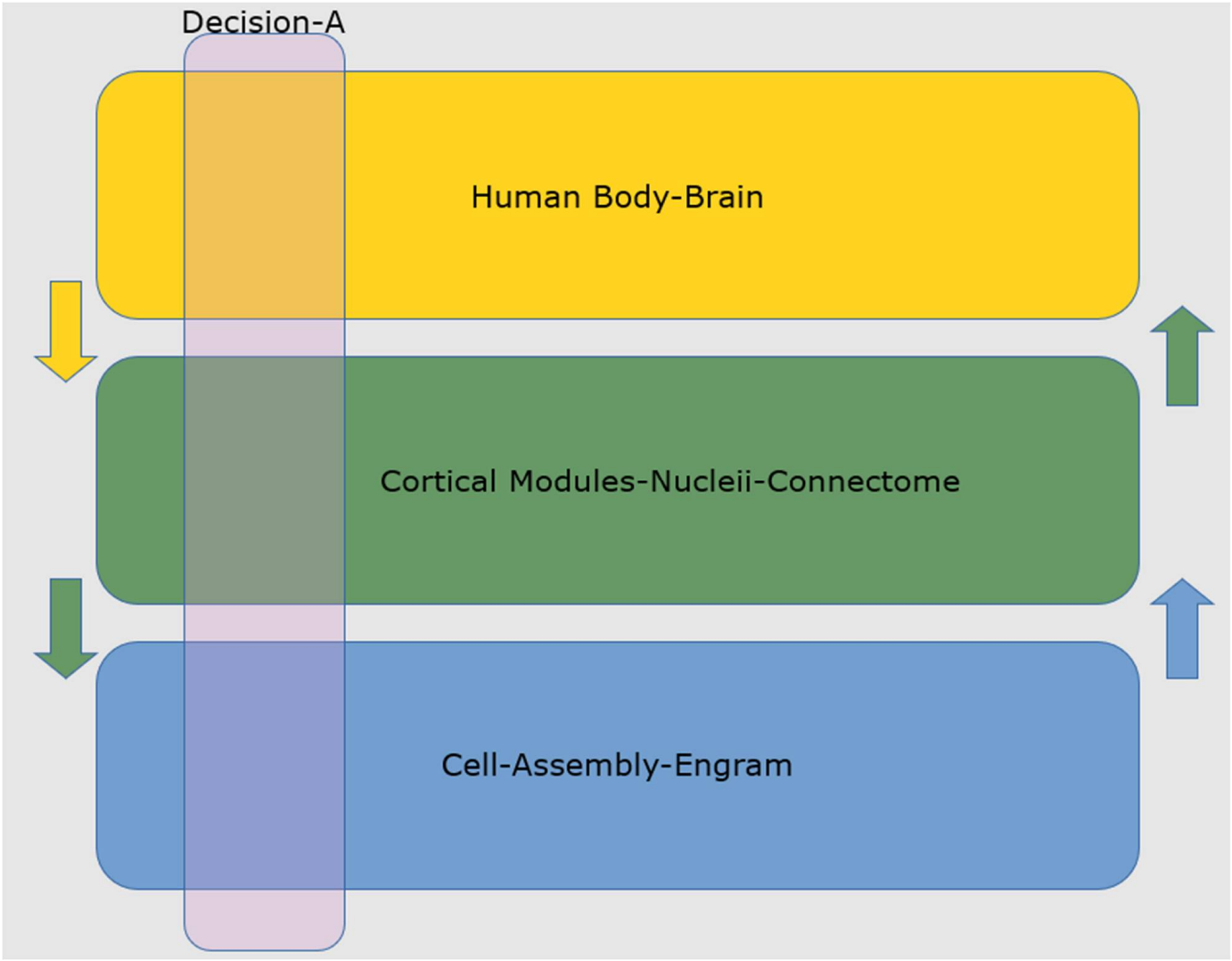

Figure 2. The Human Brain Decision Stack. Primary control flows from the bottom to the top of the stack, albeit with top-down feedback. A set of individual neurons (c. 1000 out of 10^{11}) make up a cell assembly representing a perception, concept, memory, or decision. The synchronous behavior of the cell assembly results in specific activations in brain cortical modules and nuclei via short-range and long-range axonal connections, known as the connectome. Those activations produce a cognitive response in the human body-brain, manifesting as decisions and human behaviors.

The revealing of the animal decision stack allows for a full explanation of animal decisions, as evidenced by the optogenetic studies described above [105]. By extension, such explanatory capability should be available in human subjects, provided that the spatio-temporal resolution of non-invasive functional brain imaging can be extended to the level of the human neural code (engrams). While a problem can explain an intelligence failure in the lowest level of the stack, it might also be explained by a problem above that layer. The entire functioning stack framework is necessary to explain neural intelligence failure.

While the primary flow of control in human brains is from the bottom of the stack to the top, there is an additional phenomenon that adds to the complexity and functionality of the human decision stack: feedback information and control from higher to lower levels. Such feedback is known to optimize the computational efficiency of neural computation. It has been selected for throughout the evolution of brains because of energy constraints, i.e., the human brain operates on about 20 watts of electricity, the same power as a refrigerator light [106].

5. Non-AI Machine Decision Stacks

Non-AI failures occur, often manifesting on our electronic devices. When a software engineer debugs a computer program, that process reveals the explanation for failures embedded in the program's source code. Such debugging has been evolved and engineered over the years to facilitate the central roles of human beings in debugging. While source code resembles a written language like English, it must be translated into machine code to execute on a digital computer. Machine code ultimately becomes the binary sequence of 1′s and 0′s that drive the transistors that populate the Complementary Metal–Oxide–Semiconductor (CMOS) circuitry of extant devices. Programing languages and debugging routines are human-engineered tools for explaining non-AI machine failures in digital computers. However, it is the framework of the machine decision stack illustrated in Figure 3 that enables the full explanation of such failures. Source code bugs must be compiled before they produce failure. The failure eventually manifests in the incorrect behavior of CMOS electronics.

Reviewing the machine code to derive the explanation of a program failure is an alternative to current software debugging routines. However, this would be very difficult, as evidenced by the slow speed of debugging the earliest digital computers that used machine language [107].

Since the machine code drives billions of transistors in modern computers, the *Gedanken* experiment of reading out each of their states to explain failure would be daunting. In contrast to the human brain, most explanation for failure comes from observing the system at the top, not the bottom. We observe this in a web browser that freezes on our computer desktop. The explanation is usually found in error in the source code written by a human that has propagated to the machine code level and then to transistor hardware. Hardware malfunctions are equally capable of explaining failure modes. For this reason, information flows in both directions in the machine decision stack found in modern devices, as depicted in Figure 3. For example, when the transistors inside the CPU of a modern computer become too hot, they can transmit that "distress signal" upwards to either slow down or interrupt the functional execution of a program. Furthermore, many interpreters and compilers will signal coding problems upwards in the decision stack rather than blindly writing bad machine code instructions. As a result, the arrows of information flow are bi-directional in the decision stack, as with biological brains.

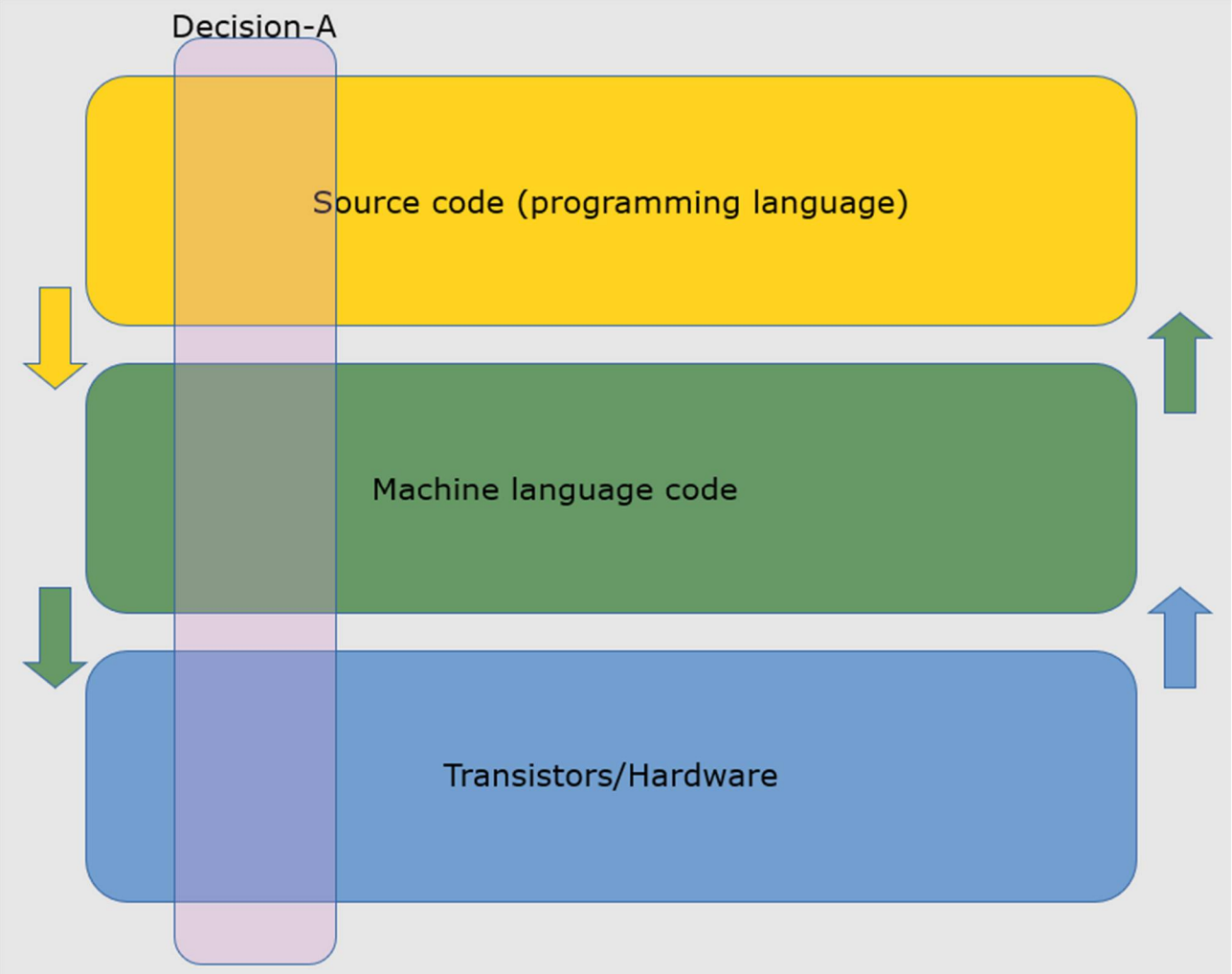

Figure 3. The decision stack for a non-intelligent digital machine. Source code for a computer program is written in a programming language (e.g., Python). The source code describes an algorithm to operate on structured data to produce a decision. For all intents and purposes, the program is deterministic. An interpreter or a compiler translates the source code into machine language, which then instructs the hardware layer to turn on and off electronic components (transistors) to produce a decision that is then translated to the human operators via an interface.

6. Explaining AI Failures: Towards an AI Decision-Stack

The challenge in providing such an explanation for AI lies in the distributed nature of virtually all ML systems. A paradigmatic example can be found in artificial neural networks where computational units, "neurons," encode a decision in their activity pattern as influenced by the weight of their respective synaptic inputs [108]. As with biological neurons in brains, the size and complexity of the network conceal which members are the "key players" in any given decision. Similar to neural systems, there are no "grandmother" cells (In neuroscience, a "grandmother cell" is a theoretical construct of the single neuron that encodes the memory of your own grandmother. Evidence suggests that grandmother cells do not exist in complex biological brains.) [109].

Additional complexity in such networks comes from the reference architecture of many AI's where heterogeneous algorithms, unstructured data storage, and a separate decision engine resides. This is schematically represented in Figure 4.

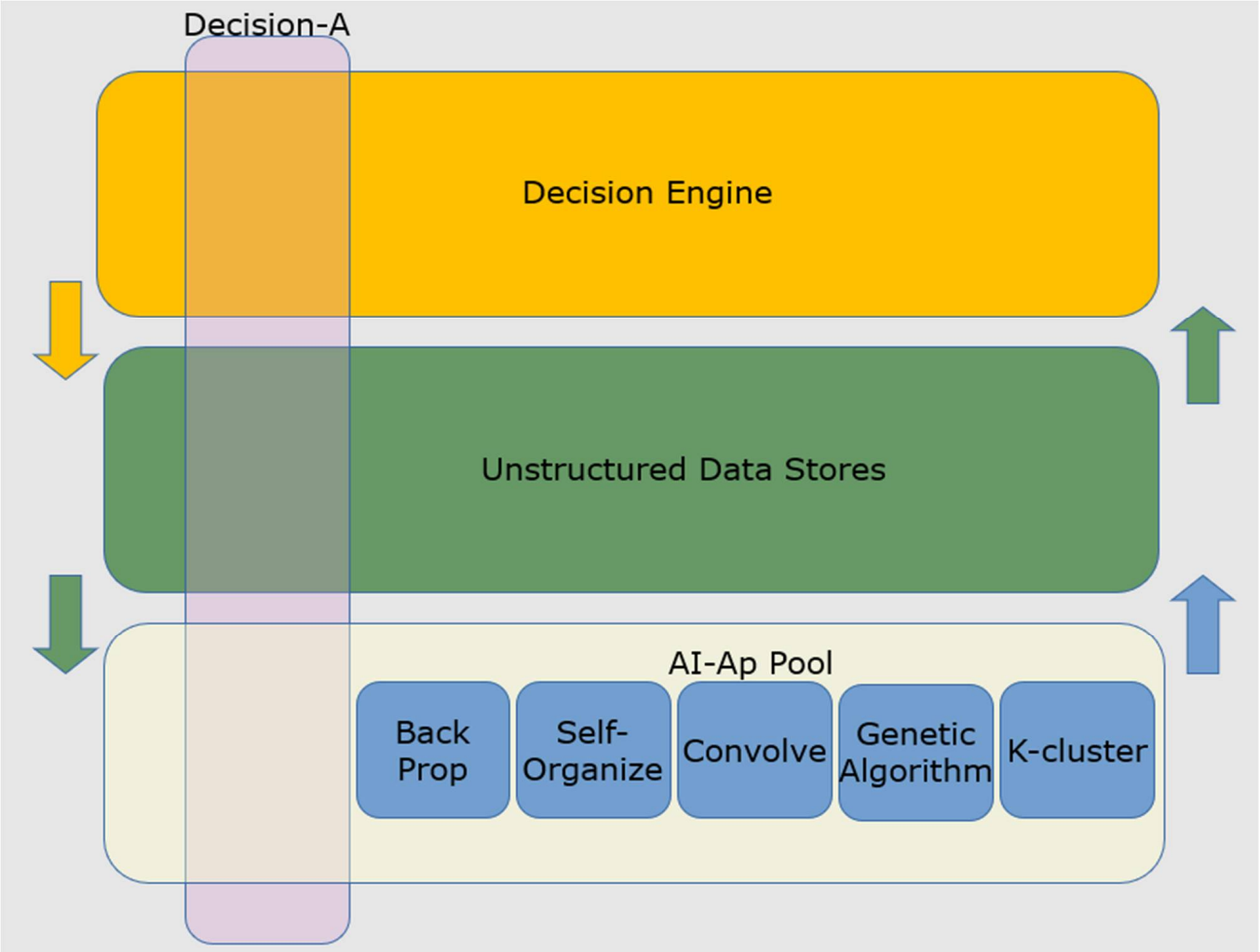

Figure 4. The decision stack for an AI decision machine. Pictured above is a schematized reference architecture upon which an AI decision stack (translucent pink) must operate. At the base are homogenous or heterogeneous populations of ML algorithms (AI-Ap Pool) that run on unstructured data. The examples in the figure include back propagation artificial neural networks, self-organizing networks, convolution algorithms, genetic algorithms, and K-clustering statistical methods. Datastores are implemented at the middle level. Such datastores, physically distributed, are accessed at the base of the reference architecture and the top as schematized by the arrows. At the top level is a decision engine that acts as a read-out of the entire decision stack. The decision engine acts as a read-out and an integral component of the decision stack that may implement some machine learning.

As with biological nervous systems, the basis for a decision may be embedded in the instantiated ML networks at the base of Figure 4. It then propagates to shape the response of the top level: an AI-decision stack. The bottom layer of the stack in Figure 4 does not represent CMOS hardware (e.g., transistors). Instead, it represents the software-defined nodes (e.g., artificial neurons) of one or several ML algorithms. In this case, it should be possible to probe the universe of these nodes for their activity during a critical decision-making process, logging that data for future analysis in a manner analogous to the neurobiology experiments described above.

Further, and crucial to explainable AI, once such an "engram" is revealed, it should be possible, in a manner analogous to biological brains, to turn off the labeled nodes and to test whether the AI's "decision" is reversed, as propagated across the AI-decision stack. Under our operational definition of explainable AI, the results of this test constitute *the explanation*, once again taking from the field of neuroscience.

7. The Neurally inspired Framework

In biological brains, an *explanation* for a single biological memory has been achieved by labeling the members of a cell assembly that, by the act of firing action potentials together, are functionally bound during memory formation (i.e., an engram). Then, by onpogenetically inactivating those cells, and only those cells, it is possible to reversibly control the recall of the specific memory [9]. We treat the mnemonic function as a specific instance of decision-making since each decision requires a corresponding memory. We introduce the notion of the decision stack, a biological "reference architecture." The members of the cell assembly are in the lowest layer of this reference architecture.

In the case of our AI framework, we describe an analogous decision stack reference architecture where the individual nodes/neurons are also at the lowest layer. The instrumentation of these nodes (analogous to functional labeling and optogenetic control) enables one to label the relevant members to test for explanation analogously. The test consists of re-running the decision with those nodes inactivated and revealing the dependence of the decision upon those specific nodes as they propagate their activity across the AI-decision stack.

It is essential to point out that, as with biological brains, individual nodes may participate in many separate decisions and that an individually flagged node may not be crucial to explaining a single decision.

8. Existing Explainable AI Methods and Our Framework

AI-based systems allow powerful predictions. However, owing to their Blackbox nature, they are not readily explainable. Existing X-AI methods make a genuine effort to break the Blackbox yet cannot fully explain all the contours of a prediction [75,110,111]. As opposed to existing X-AI methods, our decision stack framework mimics biological brain principles. The natural existence of functionally-bound neurons phylogenetically conserved across all biological brains coupled with experimentally-induced optogenetic inactivation of 'memory' cells—leading to reversable memory recall—provides an empirical basis and existence proof for our decision stacks framework. While we have already pointed out challenges, including validity and explanation issues in current X-AI methods as well as laid the ground for comparison between existing X-AI methods and our framework in Section 3, we briefly make a few more remarks here.

Applying alternative X-AI explanation methods to similar input data may lead to different results [112]. This makes comparing results from different X-AI methods a daunting task. In other words, if we do not possess prior knowledge or background information on the data producers, we are unsure which X-AI explanation is more accurate. Our framework, taking inspiration from biological brain architectures, allows for a more plausible explanation. We foresee this explanation as objective (more precise and natural) and consistent (same result for similar input data).

Existing X-AI methods also present limitations based on the data type being used [113]. Data, as we know, come in various forms: numerical, image, text, and tabular. Often such data are unstructured. In a consequential environment where users and policymakers want an explanation, a powerful X-AI method should be well-equipped to handle any kind or combination of the data. However, the existing algorithms in X-AI methods seem unprepared to handle all data types [113]. Since we base our proposed framework on biology, which can process any data to make decisions, we foresee the algorithm in our framework as advantageous.

Finally, all X-AI methods face multifaceted challenges, as extensively elaborated in a recent article [75]. Some of these challenges pertain to the dynamics of data and decisions (changing data and decisions cause different explanations) and context dependency (since outcomes may differ for individuals, general explanations for algorithms may not work). Other challenges relate to the wicked nature of the problems addressed (due to the ambiguous and poorly structured nature of the problems, the problems could warrant multiple answers as opposed to a single answer provided by current algorithms) and

contested explanations (explanations could be biased, for instance). While we do not think that our framework solves all these problems, we believe some of these problems (such as contested explanations, dynamic data, and context dependency) will likely be mitigated by the decision stacks framework that is inspired by actual biological mnemonic function in its design to return an explanation.

9. Conclusions

As the adoption of AI and ML continues to rise and reaches new audiences, increasingly complex Blackbox models (such as deep neural networks) pose explainability challenges to engineers, researchers, and policymakers alike. The future of AI's use in consequential applications in health, justice, industry, defense, and security will increasingly require explainability. After summarizing the existing X-AI methods and some of their inherent issues in this paper, we theoretically propose using an AI-decision stack framework to operationalize AI explainability analogous to modern neuroscience. The operationalization would entail instrumenting the complete ML node set and recording them all during AI decision-making. The active nodes in this process define the test set for the explanation. A successful test requires demonstrating a causal relationship between the activation of the labeled nodes and the decision.

While in this paper, we aimed to furnish a theoretical framework, future research will allow for empirical testing of this framework on actual data. In principle, X-AI methods and frameworks bridge ML systems and human systems; thus, the decision stack framework testing and further refining will need collaborations between scholars in computer science, mathematics, economics, neuroscience, behavioral psychology, public policy, and experts in human-computer interaction.

Author Contributions: Conceptualization, J.L.O., M.S.K. and N.K.; Methodology, J.L.O., M.S.K., M.N.; Validation, M.S.K., M.N., J.L.O., M.-H.L., N.K. and H.E.-A.; Formal Analysis, M.S.K., M.N. and J.L.O.; Investigation, M.S.K., M.N., N.K. and J.L.O.; Resources, N.K. and J.L.O.; Data Curation, M.S.K. and M.N.; Writing—Original Draft Preparation, M.S.K., M.N. and J.L.O.; Writing—Review & Editing, N.K., H.E.-A., M.S.K. and J.L.O.; Visualization, M.S.K., M.N., M.-H.L. and J.L.O.; Supervision, N.K. and J.L.O.; Project Administration, M.S.K.; Funding Acquisition, M.S.K., N.K., H.E.-A. and J.L.O. All authors have read and agreed to the published version of the manuscript.

Funding: Publication of this article was funded in part by the George Mason University Libraries Open Access Publishing Fund.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors express their profound thanks to Muhammad Umar Berches Niazi and anonymous reviewers for their feedback on this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gelles, D. Boeing 737 Max: What's Happened after the 2 Deadly Crashes. *The New York Times*, 22 March 2019. Available online: https://www.nytimes.com/interactive/2019/business/boeing-737-crashes.html (accessed on 13 August 2022).
2. Krichmar, J.L.; Severa, W.; Khan, M.S.; Olds, J.L. Making BREAD: Biomimetic Strategies for Artificial Intelligence Now and in the Future. *Front. Neurosci.* **2019**, *13*, 666. [CrossRef] [PubMed]
3. Cowen, T. *Average Is Over: Powering America Beyond the Age of the Great Stagnation*; Penguin Group: New York, NY, USA, 2013.
4. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484–489. [CrossRef]
5. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the Game of Go without Human Knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef]
6. Tesla Driver in Fatal "Autopilot" Crash Got Numerous Warnings: U.S. Government-Reuters. Available online: https://uk.reuters.com/article/us-tesla-crash-idUKKBN19A2XC (accessed on 21 August 2019).
7. Friedman, J.A.; Zeckhauser, R. Assessing Uncertainty in Intelligence. *Intell. Natl. Secur.* **2012**, *27*, 824–847. [CrossRef]

8. More Efficient Machine Learning Could Upend the AI Paradigm. MIT Technology Review. Available online: https://www.technologyreview.com/2018/02/02/145844/more-efficient-machine-learning-could-upend-the-ai-paradigm/ (accessed on 13 August 2022).
9. Liu, X.; Ramirez, S.; Pang, P.T.; Puryear, C.B.; Govindarajan, A.; Deisseroth, K.; Tonegawa, S. Optogenetic Stimulation of a Hippocampal Engram Activates Fear Memory Recall. *Nature* **2012**, *484*, 381–385. [CrossRef] [PubMed]
10. Fellows, L.K. The Neuroscience of Human Decision-Making Through the Lens of Learning and Memory. *Curr. Top. Behav. Neurosci.* **2018**, *37*, 231–251. [CrossRef]
11. Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
12. Wallace, N.; Castro, D. The Impact of the EU's New Data Protection Regulation on AI. Center for Data Innovation. 2018. Available online: https://www2.datainnovation.org/2018-impact-gdpr-ai.pdf (accessed on 27 March 2018).
13. Goodman, B.; Flaxman, S. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Mag.* **2017**, *38*, 50–57. [CrossRef]
14. Rossini, P.M.; Burke, D.; Chen, R.; Cohen, L.G.; Daskalakis, Z.; Di Iorio, R.; Di Lazzaro, V.; Ferreri, F.; Fitzgerald, P.B.; George, M.S.; et al. Non-Invasive Electrical and Magnetic Stimulation of the Brain, Spinal Cord, Roots and Peripheral Nerves: Basic Principles and Procedures for Routine Clinical and Research Application. An Updated Report from an I.F.C.N. Committee. *Clin. Neurophysiol.* **2015**, *126*, 1071–1107. [CrossRef]
15. Yibo, C.; Hou, K.; Zhou, H.; Shi, H.; Liu, X.; Diao, X.; Ding, H.; Li, J.; de Vaulx, C. 6LoWPAN Stacks: A Survey. In Proceedings of the 2011 7th International Conference on Wireless Communications, Networking and Mobile Computing, Wuhan, China, 23–25 September 2011; pp. 1–4. [CrossRef]
16. Clancey, W.J. From GUIDON to NEOMYCIN and HERACLES in Twenty Short Lessons. *AI Mag.* **1986**, *7*, 40. [CrossRef]
17. Moore, J.D.; Swartout, W.R. Pointing: A Way Toward Explanation Dialogue. In Proceedings of the 8th National Conference on Artificial Intelligence, Boston, MA, USA, 29 July–3 August 1990; Shrobe, H.E., Dietterich, T.G., Swartout, W.R., Eds.; AAAI Press; The MIT Press: Cambridge, MA, USA, 1990; Volume 2, pp. 457–464.
18. Biran, O.; Cotton, C. Explanation and Justification in Machine Learning: A Survey. In Proceedings of the IJCAI-17 Workshop on Explainable AI (XAI), Melbourne, Australia, 20 August 2017; Volume 8, pp. 8–13.
19. Paudyal, P. Should AI Explain Itself? Or Should We Design Explainable AI so that It Doesn't Have to. Medium. Available online: https://towardsdatascience.com/should-ai-explain-itself-or-should-we-design-explainable-ai-so-that-it-doesnt-have-to-90e75bb6089e (accessed on 13 August 2022).
20. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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, KDD '16, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1135–1144. [CrossRef]
21. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias. ProPublica. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=Tu5C70R2pCBv8Yj33AkMh2E-mHz3d6iu (accessed on 13 August 2022).
22. Kirsch, A. Explain to Whom? Putting the User in the Center of Explainable AI. In Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 Co-Located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy, 14 November 2017.
23. Mueller, S.T.; Hoffman, R.R.; Clancey, W.; Emrey, A.; Klein, G. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *arXiv* **2019**. [CrossRef]
24. Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Scotland, UK, 4–9 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–15. [CrossRef]
25. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects. *arXiv* **2019**. [CrossRef]
26. Sheh, R.; Monteath, I. Defining Explainable AI for Requirements Analysis. *Künstl Intell.* **2018**, *32*, 261–266. [CrossRef]
27. Kim, T.W. Explainable Artificial Intelligence (XAI), the Goodness Criteria and the Grasp-Ability Test. *arXiv* **2018**. [CrossRef]
28. Adler, P.; Falk, C.; Friedler, S.A.; Nix, T.; Rybeck, G.; Scheidegger, C.; Smith, B.; Venkatasubramanian, S. Auditing Black-Box Models for Indirect Influence. *Knowl. Inf. Syst.* **2018**, *54*, 95–122. [CrossRef]
29. Wachter, S.; Mittelstadt, B.; Floridi, L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *Int. Data Priv. Law* **2017**, *7*, 76–99. [CrossRef]
30. Hayes, B.; Shah, J.A. Improving Robot Controller Transparency Through Autonomous Policy Explanation. In Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI '17, Vienna, Austria, 6–9 March 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 303–312. [CrossRef]
31. Fallon, C.K.; Blaha, L.M. Improving Automation Transparency: Addressing Some of Machine Learning's Unique Challenges. In *Augmented Cognition: Intelligent Technologies*; Schmorrow, D.D., Fidopiastis, C.M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; pp. 245–254. [CrossRef]
32. What Does It Mean to Ask for an "Explainable" Algorithm? Available online: https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/ (accessed on 14 August 2022).
33. Bengio, Y.; Lecun, Y.; Hinton, G. Deep Learning for AI. Commun. *ACM* **2021**, *64*, 58–65. [CrossRef]

34. Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; Zhu, J. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In *Natural Language Processing and Chinese Computing*; Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; pp. 563–574. [CrossRef]
35. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI Methods—A Brief Overview. In *xxAI-Beyond Explainable AI, Proceedings of the International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Vienna, Austria, 18 July 2020*; Revised and Extended, Papers; Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2022; pp. 13–38. [CrossRef]
36. Lipton, Z.C. The Mythos of Model Interpretability. *arXiv* **2017**. [CrossRef]
37. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable Artificial Intelligence: A Survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 0210–0215. [CrossRef]
38. Madsen, A.; Reddy, S.; Chandar, S. Post-Hoc Interpretability for Neural NLP: A Survey. *ACM Comput. Surv.* **2022**. [CrossRef]
39. Zhang, Y.; Weng, Y.; Lund, J. Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. *Diagnostics* **2022**, *12*, 237. [CrossRef]
40. Yera, R.; Alzahrani, A.A.; Martínez, L. Exploring Post-Hoc Agnostic Models for Explainable Cooking Recipe Recommendations. *Knowl. Based Syst.* **2022**, *251*, 109216. [CrossRef]
41. Pintelas, E.; Livieris, I.E.; Pintelas, P. A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability. *Algorithms* **2020**, *13*, 17. [CrossRef]
42. Grau, I.; Sengupta, D.; Lorenzo, M.M.G. Grey-Box Model: An Ensemble Approach for Addressing Semi-Supervised Classification Problems. 2016. Available online: https://kulak.kuleuven.be/benelearn/papers/Benelearn_2016_paper_45.pdf (accessed on 13 August 2022).
43. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses from a Practical Point of View. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
44. Taherkhani, N.; Sepehri, M.M.; Khasha, R.; Shafaghi, S. Ranking Patients on the Kidney Transplant Waiting List Based on Fuzzy Inference System. *BMC Nephrol.* **2022**, *23*, 31. [CrossRef] [PubMed]
45. Zaitseva, E.; Levashenko, V.; Rabcan, J.; Krsak, E. Application of the Structure Function in the Evaluation of the Human Factor in Healthcare. *Symmetry* **2020**, *12*, 93. [CrossRef]
46. Nazemi, A.; Fatemi Pour, F.; Heidenreich, K.; Fabozzi, F.J. Fuzzy Decision Fusion Approach for Loss-given-Default Modeling. *Eur. J. Oper. Res.* **2017**, *262*, 780–791. [CrossRef]
47. Zaitseva, E.; Levashenko, V.; Kvassay, M.; Deserno, T.M. Reliability Estimation of Healthcare Systems Using Fuzzy Decision Trees. In Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, Gdansk, Poland, 11–14 September 2016; pp. 331–340. [CrossRef]
48. Dubois, D.; Prade, H. Fuzzy Set and Possibility Theory-Based Methods in Artificial Intelligence. *Artif. Intell.* **2003**, *148*, 1–9. [CrossRef]
49. Garibaldi, J.M. The Need for Fuzzy AI. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 610–622. [CrossRef]
50. Robnik-Šikonja, M.; Kononenko, I. Explaining Classifications for Individual Instances. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 589–600. [CrossRef]
51. Zhang, Y.; Xu, F.; Zou, J.; Petrosian, O.L.; Krinkin, K.V. XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction. In Proceedings of the 2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT), Saint Petersburg, Russia, 16 June 2021; 2021; pp. 13–16. [CrossRef]
52. Aliramezani, M.; Norouzi, A.; Koch, C.R. A Grey-Box Machine Learning Based Model of an Electrochemical Gas Sensor. *Sens. Actuators B Chem.* **2020**, *321*, 128414. [CrossRef]
53. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
54. Machlev, R.; Heistrene, L.; Perl, M.; Levy, K.Y.; Belikov, J.; Mannor, S.; Levron, Y. Explainable Artificial Intelligence (XAI) Techniques for Energy and Power Systems: Review, Challenges and Opportunities. *Energy AI* **2022**, *9*, 100169. [CrossRef]
55. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems—NIPS'17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
56. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *arXiv* **2014**. [CrossRef]
57. Danesh, T.; Ouaret, R.; Floquet, P.; Negny, S. Interpretability of Neural Networks Predictions Using Accumulated Local Effects as a Model-Agnostic Method. In *Computer Aided Chemical Engineering*; Montastruc, L., Negny, S., Eds.; 32 European Symposium on Computer Aided Process Engineering; Elsevier: Amsterdam, The Netherlands, 2022; Volume 51, pp. 1501–1506. [CrossRef]
58. Belle, V.; Papantonis, I. Principles and Practice of Explainable Machine Learning. *Front. Big Data* **2021**, *4*, 688969. [CrossRef] [PubMed]
59. Sairam, S.; Srinivasan, S.; Marafioti, G.; Subathra, B.; Mathisen, G.; Bekiroglu, K. Explainable Incipient Fault Detection Systems for Photovoltaic Panels. *arXiv* **2020**. [CrossRef]

60. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Trans. Syst. Man Cybern. Part C* **2012**, *42*, 463–484. [CrossRef]
61. Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Gradient-Based Attribution Methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; pp. 169–191. [CrossRef]
62. Lei, T.; Barzilay, R.; Jaakkola, T. Rationalizing Neural Predictions. *arXiv* **2016**. [CrossRef]
63. Jain, S.; Wiegreffe, S.; Pinter, Y.; Wallace, B.C. Learning to Faithfully Rationalize by Construction. *arXiv* **2020**. [CrossRef]
64. Esmaeili, B.; Wu, H.; Jain, S.; Bozkurt, A.; Siddharth, N.; Paige, B.; Brooks, D.H.; Dy, J.; Meent, J.-W. Structured Disentangled Representations. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, Okinawa, Japan, 16–18 April 2019; Proceedings of Machine Learning Research; pp. 2525–2534.
65. Palczewska, A.; Palczewski, J.; Marchese Robinson, R.; Neagu, D. Interpreting Random Forest Classification Models Using a Feature Contribution Method. In *Integration of Reusable Systems*; Bouabana-Tebibel, T., Rubin, S.H., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2014; pp. 193–218. [CrossRef]
66. Tolomei, G.; Silvestri, F.; Haines, A.; Lalmas, M. Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 465–474. [CrossRef]
67. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*. [CrossRef]
68. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; Sayres, R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *arXiv* **2018**. [CrossRef]
69. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv* **2016**. [CrossRef]
70. Vilone, G.; Longo, L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 615–661. [CrossRef]
71. Welling, S.H.; Refsgaard, H.H.F.; Brockhoff, P.B.; Clemmensen, L.H. Forest Floor Visualizations of Random Forests. *arXiv* **2016**. [CrossRef]
72. Wongsuphasawat, K.; Smilkov, D.; Wexler, J.; Wilson, J.; Mané, D.; Fritz, D.; Krishnan, D.; Viégas, F.B.; Wattenberg, M. Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 1–12. [CrossRef] [PubMed]
73. Hendricks, L.A.; Hu, R.; Darrell, T.; Akata, Z. Grounding Visual Explanations. *arXiv* **2018**. [CrossRef]
74. Pawelczyk, M.; Haug, J.; Broelemann, K.; Kasneci, G. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 3126–3132. [CrossRef]
75. de Bruijn, H.; Warnier, M.; Janssen, M. The Perils and Pitfalls of Explainable AI: Strategies for Explaining Algorithmic Decision-Making. *Gov. Inf. Q.* **2022**, *39*, 101666. [CrossRef]
76. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
77. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]
78. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
79. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and Explainability of Artificial Intelligence in Medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef]
80. Alcorn, M.A.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; Ku, W.-S.; Nguyen, A. Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. *arXiv* **2019**. [CrossRef]
81. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1625–1634. [CrossRef]
82. Su, J.; Vargas, D.V.; Kouichi, S. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Trans. Evol. Computat.* **2019**, *23*, 828–841. [CrossRef]
83. Fatal Tesla Self-Driving Car Crash Reminds Us That Robots Aren't Perfect. IEEE Spectrum. Available online: https://spectrum.ieee.org/fatal-tesla-autopilot-crash-reminds-us-that-robots-arent-perfect (accessed on 13 August 2022).
84. Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients-IEEE Spectrum. Available online: https://spectrum.ieee.org/racial-bias-found-in-algorithms-that-determine-health-care-for-millions-of-patients (accessed on 13 August 2022).
85. Newman, J. *Explainability Won't Save AI*; Brookings: Washington, DC, USA, 2021.
86. Bansal, N.; Agarwal, C.; Nguyen, A. SAM: The Sensitivity of Attribution Methods to Hyperparameters 2020. Poster. Available online: https://bnaman50.github.io/SAM/CVPR_2020_SAM_Poster.pdf (accessed on 13 August 2022).
87. Explainable AI Won't Deliver. Here's Why. | HackerNoon. Available online: https://hackernoon.com/explainable-ai-wont-deliver-here-s-why-6738f54216be (accessed on 14 August 2022).

88. 7 Revealing Ways AIs Fail. IEEE Spectrum. Available online: https://spectrum.ieee.org/ai-failures (accessed on 14 August 2022).
89. Medical Imaging AI Software Is Vulnerable to Covert Attacks. IEEE Spectrum. Available online: https://spectrum.ieee.org/medical-imaging-ai-software-vulnerable-to-covert-attacks (accessed on 14 August 2022).
90. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. *arXiv* **2014**. [CrossRef]
91. Nantais, J. Does Your Data Science Project Actually Do What You Think It Does? Medium. Available online: https://towardsdatascience.com/internal-validity-in-data-science-c44c1a2f194f (accessed on 14 August 2022).
92. Bansal, N.; Agarwal, C.; Nguyen, A. SAM: The Sensitivity of Attribution Methods to Hyperparameters. *arXiv* **2020**. [CrossRef]
93. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: Removing Noise by Adding Noise. *arXiv* **2017**. [CrossRef]
94. Olds, J. *Ideas Lab for Imagining Artificial Intelligence and Augmented Cognition in the USAF of 2030*; George Mason University: Arlington, VA, USA, 2019; p. 134. Available online: https://apps.dtic.mil/sti/pdfs/AD1096469.pdf (accessed on 22 August 2019).
95. Vosskuhl, J.; Strüber, D.; Herrmann, C.S. Non-Invasive Brain Stimulation: A Paradigm Shift in Understanding Brain Oscillations. *Front. Hum. Neurosci.* **2018**, *12*, 211. [CrossRef]
96. Zrenner, C.; Belardinelli, P.; Müller-Dahlhaus, F.; Ziemann, U. Closed-Loop Neuroscience and Non-Invasive Brain Stimulation: A Tale of Two Loops. *Front. Cell. Neurosci.* **2016**, *10*, 92. [CrossRef] [PubMed]
97. Pollok, B.; Boysen, A.-C.; Krause, V. The Effect of Transcranial Alternating Current Stimulation (TACS) at Alpha and Beta Frequency on Motor Learning. *Behav. Brain Res.* **2015**, *293*, 234–240. [CrossRef]
98. Antal, A.; Nitsche, M.A.; Kincses, T.Z.; Kruse, W.; Hoffmann, K.-P.; Paulus, W. Facilitation of Visuo-Motor Learning by Transcranial Direct Current Stimulation of the Motor and Extrastriate Visual Areas in Humans. *Eur. J. Neurosci.* **2004**, *19*, 2888–2892. [CrossRef]
99. Zaehle, T.; Sandmann, P.; Thorne, J.D.; Jäncke, L.; Herrmann, C.S. Transcranial Direct Current Stimulation of the Prefrontal Cortex Modulates Working Memory Performance: Combined Behavioural and Electrophysiological Evidence. *BMC Neurosci.* **2011**, *12*, 2. [CrossRef]
100. Bartolotti, J.; Bradley, K.; Hernandez, A.E.; Marian, V. Neural Signatures of Second Language Learning and Control. *Neuropsychologia* **2017**, *98*, 130–138. [CrossRef] [PubMed]
101. Anggraini, D.; Glasauer, S.; Wunderlich, K. Neural Signatures of Reinforcement Learning Correlate with Strategy Adoption during Spatial Navigation. *Sci. Rep.* **2018**, *8*, 10110. [CrossRef] [PubMed]
102. Heekeren, H.R.; Marrett, S.; Bandettini, P.A.; Ungerleider, L.G. A General Mechanism for Perceptual Decision-Making in the Human Brain. *Nature* **2004**, *431*, 859–862. [CrossRef] [PubMed]
103. Hsu, M.; Bhatt, M.; Adolphs, R.; Tranel, D.; Camerer, C.F. Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making. *Science* **2005**, *310*, 1680–1683. [CrossRef]
104. Dudai, Y. The Neurobiology of Consolidations, or, How Stable Is the Engram? *Annu. Rev. Psychol.* **2004**, *55*, 51–86. [CrossRef]
105. Ramirez, S.; Liu, X.; Lin, P.-A.; Suh, J.; Pignatelli, M.; Redondo, R.L.; Ryan, T.J.; Tonegawa, S. Creating a False Memory in the Hippocampus. *Science* **2013**, *341*, 387. [CrossRef]
106. Sandberg, A. Energetics of the Brain and AI. *arXiv* **2016**. [CrossRef]
107. Debugging Tools for High Level Languages-Satterthwaite-1972-Software: Practice and Experience-Wiley Online Library. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380020303 (accessed on 14 August 2022).
108. Yao, X. Evolving Artificial Neural Networks. *Proc. IEEE* **1999**, *87*, 1423–1447. [CrossRef]
109. Bowers, J.S. On the Biological Plausibility of Grandmother Cells: Implications for Neural Network Theories in Psychology and Neuroscience. *Psychol. Rev.* **2009**, *116*, 220–251. [CrossRef]
110. Janssen, M.; Kuk, G. The Challenges and Limits of Big Data Algorithms in Technocratic Governance. *Gov. Inf. Q.* **2016**, *33*, 371–377. [CrossRef]
111. Wan, A. What Explainable AI Fails to Explain (and How We Fix that). Medium. Available online: https://towardsdatascience.com/what-explainable-ai-fails-to-explain-and-how-we-fix-that-1e35e37bee07 (accessed on 15 August 2022).
112. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
113. Joshi, K. Deep Dive into Explainable AI: Current Methods and Challenges. Arya-xAI 2022. Available online: https://medium.com/arya-xai/deep-dive-into-explainable-ai-current-methods-and-challenges-2e9912d73136 (accessed on 1 February 2022).