philosophies

MDPI

Article

On Explainable AI and Abductive Inference

Kyrylo Medianovskyi [1,†] and Ahti-Veikko Pietarinen [2,*,†]

[1] Department of Software Science, School of Information Technologies, Tallinn University of Technology, 19086 Tallinn, Estonia; kyrylo.medianovskyi@taltech.ee
[2] Ragnar Nurkse Department of Innovation and Governance, School of Business and Governance, Tallinn University of Technology, 19086 Tallinn, Estonia
* Correspondence: ahti-veikko.pietarinen@taltech.ee
† These authors contributed equally to this work.

Abstract: Modern explainable AI (XAI) methods remain far from providing human-like answers to 'why' questions, let alone those that satisfactorily agree with human-level understanding. Instead, the results that such methods provide boil down to sets of causal attributions. Currently, the choice of accepted attributions rests largely, if not solely, on the explainee's understanding of the quality of explanations. The paper argues that such decisions may be transferred from a human to an XAI agent, provided that its machine-learning (ML) algorithms perform genuinely abductive inferences. The paper outlines the key predicament in the current inductive paradigm of ML and the associated XAI techniques, and sketches the desiderata for a truly participatory, second-generation XAI, which is endowed with abduction.

Keywords: explainable AI (XAI); machine learning; abduction; induction; explanation; understanding; causal attributions; counterfactuals

Citation: Medianovskyi, K.; Pietarinen, A.-V. On Explainable AI and Abductive Inference. *Philosophies* **2022**, *7*, 35. https://doi.org/10.3390/philosophies7020035

Academic Editor: Woosuk Park

Received: 14 February 2022
Accepted: 19 March 2022
Published: 23 March 2022

1. Introduction

The paradigm of machine learning (ML)—indeed in the nearly Kuhnian sense of the term as the accepted state of normal science—is to perform reasoning by *induction*. The success of the software code in which deep learning algorithms—supervised, unsupervised, semi-supervised or self-supervised alike—have been written has, for a long time, hinged on the assumption that the kind of reasoning from big data that the code has been implemented to perform has to take place inductively by finding out those presumed rules and models that otherwise would not have been discovered, in terms of statistical generalisations from specific and large-enough sets of unclassified data that qualify as such fair samples as the basis of those generalisations.

Therefore, also the validity of the discoveries of a great variety of possible relationships (real and non-real, positive and negative, etc.) that are to give rise to successful pattern recognition and matching, and the classification and detection of law-like dependencies, or even those of predictions, ensue to a significant degree from the intentions of the system's modellers, their fallible decisions concerning the data selection, and even from the multiplicity of biases and noise that characterises human-led operations in the science of artificial intelligence (AI).

This nearly Baconian definition of (weak) induction undergirds recent and current research in ML. It therefore also—though often unintentionally—constrains the directions that various methods for software development in AI could take in their relevant theoretical subdomains. Our wider point is to propose a way to spark a deliberative effect on the automatised discoveries that emancipate the field from the trenches of gradient learning from data, and thus from the associated human idiosyncracies that contribute to the initial pseudo-weightings of nodes of deep neural nets that might succeed in emulating the workings of synaptic connections and their complex and largely hidden interactions.

Although the limitations of purely inductive modes of inferring and reasoning were vocally pointed out over two decades ago in contexts such as diagnosis, planning and design [1], those voices have remained largely ineffectual in resulting in the kinds of transformative large-scale changes that are indispensable in order for these fields to move from (narrow, industry-driven) AI solutions toward truly novel and cross-disciplinary strong, artificial general intelligence (AGI) development. It is the latter realm in which the inductive paradigm meets the end of the road. Instead of broadening the theoretical base of the types of reasoning, ML has stuck to its guns of induction and founded even the otherwise promising newcomer of self-supervised learning on just the same modalities of inductive reasoning.

In the present paper, we propose that this inadequacy presented by the narrow focussing on only one type of ampliative inference risks alike the paradigmatisation of fields closely related to ML. In particular, we have in mind explainable artificial intelligence (XAI). If XAI was to follow in the footsteps of ML, we fear that it would preclude XAI from reaching the stated goals of that important programme, namely to automatise the discovery and formulation of explanations about what is going on in complex AI systems. For such explanations, it would have to be given in terms of not only explaining the mechanics of the systems in some simplified technical sense, but also in terms of the optics of how the workings of various AI-led instrumentations are to be perceived, comprehended and ultimately understood by their users and consumers across various contexts, cultures and circumstances.

To mitigate the potential negative impact of these worries, the paper argues that decisions to choose accepted attributions that currently rest on the explainee's understanding of the quality of explanations, may be translated from a human to an XAI agent provided that its novel ML algorithms can perform genuinely *abductive* inferences. We delve into the details of how to explain the meaning of abduction in due course. For now, the motivation for our treatise comes from the observation that, taking human explanations as the source and XAI, the target rests on the assumption that abduction is that precious faculty of entities we call minds (though not exclusively human minds) which have emerged and evolved in affinity and interaction with nature over evolutionary time. While some such abductions may be elevated to the status of explanations, not all of them need to do so. Nevertheless, it is by taking the more scenic route that traverses through abduction, one that is not rushing to shortcut into induction, that one can hope to appreciate the complexities and influences that are involved in one's explanation-seeking activities, be they of human or machine origin alike.

We begin by outlining the key predicament in the current inductive paradigm of ML and the associated XAI techniques, and then, in the latter sections, return to the issue of abduction and sketch the desiderata for a truly participatory, second-generation XAI endowed with abduction.

2. The Many Purposes of XAI

XAI has set itself an ambitious goal of making autonomous AI systems succeed in responding to requests for explanations of its own states, behaviours and outputs. The need for XAI has arisen from perceiving current ML as opaque or even representing solely behaviouristic black-box learning models that are in some sense incomprehensible and ill-motivated in their actions and intentions [2]. The degree of opacity of relevant instrumentations and codes may range from a total intractability of computations to the generation of an unlimited number of suggested models and rules serving as possible explanations, the preferences between which are nowhere explicitly stated or even defined. Given the vast range of explanations, from incoherent to completely unsatisfactory, the risk of unwanted behaviour and individual, social, economic and psychological harm following the unintended consequences of opaque algorithmic implementations is greatly accelerated.

Among the usual shortcomings of the proposed explanation techniques delivered to the end users are the following:

1. Even if the output of the XAI system (agent) consists of *reasonable* (i.e., distinguishable from noise) *causal attributions* of its behaviour that emanate from the underlying data-sets, the system may be incapable of automatically drawing definite conclusions, or converging to a reasonably small number of satisfactory conclusions, to the all-important 'why' question:
    - Why were some particular things or events that the model predicts produced by the system that is intended to perform causal explanations?
2. In general, predictions are not explanations. Even if the system manages to generate a limited set of conclusions to that 'why' question, the residue is as follows:
    - Why would any of those conclusions bring it about that the explainee *understands* some of those conclusions, or, in the very least, is reasonably and acceptably *satisfied* by the conclusion offered (say, because of some unfettered generic trust in algorithmic decision making)?

The wider point lurking here is that, conceptually, explanation is not understanding. Von Wright [3] famously calls causal explanations 'Galileian', in contradistinction to 'Aristotelian' explanations that are conceived in terms of purposes and intentions to act in certain ways in certain circumstances. The Aristotelian view calls for an *interpretation* of some kind that the explainee assigns to the proposed explanation. In other words, the two questions above boil down to no less than what the semantics of what 'why' questions might look like, and to the resulting problem of what the requests for information to the system that the explainee desires to put forward should be and in what form. In general, one might ask, just what kind of human–machine interaction is required to be present in a given technological context, for an explanation to translate into understanding in the presence of the variety of circumstances and cultures surrounding the ideas of meaning and technological progress?[1]

The theoretical issues affecting XAI thus reflect long-standing issues in the philosophy of language, science and communication. The important hermeneutics between explanations and understanding, and their differing roles and traditions in science has been studied since Ref. [3] and beyond. The semantics of 'why' questions has been the focus of logic and semantics, at least since Hintikka's seminal work [4], and countless others have joined the choir. Conclusiveness conditions of answers provided to requests for information were thoroughly analysed in the context of presuppositions and epistemic logic, beginning with [5], among others.

The meaning of 'why' questions continues to define a core research area in linguistic semantics and pragmatics, where it is well recognised how highly sensitive to misunderstanding and misinterpretation such questions are, given their dependence on the variety of speech, dialect, emphasis, stress and other prosodic and contextual factors as well as the utterer's intentions and background beliefs. It is no exaggeration to describe the situation in XAI as a research programme that needs, if successful, to accommodate a whole host of recognised problems in the fundamental science of the mind, involving the vast research on what the interconnected workings of language, communication and theories of logic and reasoning may look like.

Given the issues at stake, it is the latter path that we propose to explore a little further in the present paper. Following the rebirth of research on explainability and its characterisations in the AI domain, it is important to examine what comes out of the exploration of the intersecting areas of human and machine learning and discovery. Our main claim is that the mode of reasoning that should be taken primarily to nudge the field of XAI forward is not induction but abduction. Abduction was first proposed by Charles S. Peirce [6] as the first of three, not two, and essentially distinct kinds, of reasoning: along with deduction and induction, he isolated a third form, which he variously called "hypothesis", "abduction", "presumption", and "retroduction". Early on, Peirce conceived of induction and abduction as inversions of the syllogism. Later, he would abandon the syllogistic framework of the early theory and construe abduction and the method of finding

and selecting explanatory hypotheses; in the mature theory of abduction, it is not just another kind of reasoning along deduction and induction but also a distinct step in the overall process of scientific inquiry, whose typical pattern is from abduction to deduction to induction.

As the inductive paradigm is clearly inadequate to accomplish these full moves and loops, the field of ML should follow suit and seek better ways out of its solitary paradigm of inductive reasoning. One expects ML to be able to expand its inferential horizons in order to not only work better with big data at large, but to accommodate all relevant information that characterises human intentional behaviour and communication. As with most scientific information and evidence, information that is fundamentally uncertain, noisy, uni- as opposed to multi-dimensional (and hence potentially biased), anecdotal, practice-based, non-theory-laden and in many, if not most, instances, in need of interpretation and constant and considered interventions may not be of much use in automatised decision making. It is in such fundamentally uncertain and surprising problematic circumstances encountered by semi-to-full autonomous systems, in a toddler-like need of human intervention and guidance in order to even begin engaging in meaningful interaction with their environment, that prospects of abduction may shine above induction and deduction. For although induction can handle measurable uncertainty and explicit inconsistency, deduction cannot; and although abduction can handle fundamental uncertainty in which the problem space is ill-defined and non-measurable, induction cannot.

First, we look into some proposed and prospective methods of XAI and what could be done with them. We propose that one would do well in incorporating abductive reasoning into the research context of artificial agents in contemporary XAI, motivated by our observation of how such an account of explanations in the fully abductive sense of the term can much improve the state of XAI at its conceptual and design levels. We begin with some examples of the current idea of what the methods of XAI might be.

3. Methods of Explainable AI

3.1. Post-Hoc Explainability

How to automate the generation of explanatory processes? First, let us consider some *post-hoc* explainability techniques, which may be applicable for certain complex and practically opaque ML models, prototypically those arising from deep neural network architectures.

For simplicity, we highlight just two existing classes of XAI methods: *attribution of feature importance* (AFI) and *counterfactual examples* (CF). Feature importance techniques provide a heat-map (i.e., saliency map) for an imagery data set, or they attribute weights to tabular input features. Counterfactual examples target small changes made in the input data point to have the model predict a different outcome.

3.1.1. Feature Importance

Currently, one cannot provide a reliable algorithmic connection between a saliency map and a summary text that would make enough sense for an expert explainee. That is, it is an unsolved problem as to how to generate meaningful and succinct linguistic output from given data visualisation techniques that encode visually important features of a topographical chart. For optimal prediction accuracy, such algorithms might do well if they were to mimic the successful mechanisms of how attention is formed in ventral and dorsal visual pathways that process, respectively, the perspectives and identities of the objects of the data in question. The two pathways also interact for optimal prediction about both the location and identity of objects; the general lesson is that the two pathways are reflected also in the semantics of natural language in terms of "knowing what" and "knowing where".

There are possible wrappings, such as linguistic protoforms, but those only provide containers or placeholders to what needs to be not only a syntactically correct and humanly comprehensible string of words, but also one that is a semantically and pragmatically

sensible, commonplace and useful piece of language. Self-supervised learning methods may do relatively well in efficiently filling out those blanks to yield grammatical outputs; the crux of the matter nonetheless is that semantics does not follow syntax, and that all three—syntax, semantics and pragmatics of language—are interconnected already at the level of predicate (feature) formation. A much-debated case in theoretical linguistics comes from Recanati [7], for example, which argues that notions such as time, location, person, circumstance, context, etc. are not arguments of the verb although they affect the truth conditions of the propositions in which predicates and relations appear.

Other models of feature attribution might allow the expert to make an inference based on their background knowledge and some elements of common sense, but building an automatised method that returns, from the reasoning, the same conclusions as an expert might be an intractable problem. The problem of intractability is heightened when the data limitations of the model are reached, and in order for one to proceed with the inference, one needs to draw from elements of experience not encoded in the search space at all, or those elements are unretrievable because the features (although encoded in some way), the structure of the search space, and the possible continuations of its branches are fundamentally uncertain to the reasoner.

For these reasons, and despite the fact that there are internal causes that are in a feature space of a model somewhere, there are also abundant collections of external causes that are not directly added to the generative model as its random input variables. As those external and contextual causes often do produce significant effects as to what counts as the relevant and important features of the saliency map, any automatised conversion of what the sum effect of those features of the map is gong to be into a sensible piece of language that would count as explanation presents colossal, if not insurmountable, challenges.

3.1.2. Counterfactual Examples

The outcome of the counterfactual (CF) technique is a production of causal attributions. That method needs strict constraints in order to prevent false (or impossible, irrelevant, inapplicable, and severely biased) examples. In ML, CF is expected to typify the kinds of generic but example-laden explanations that give the explainees what they seem to be seeking in explanations: small variations in the input data that would or could merit a reassessment or retraction of the actual proposed action [8,9].

The idea of CF and the reasoning that it represents is a well-studied problem in the logic and philosophy of language [10–12], and it is a welcome addition in the toolkit of ML and XAI research to probe possible applications. The outcomes of the CF technique could be fully relevant to the inner workings of the model, and they even might align quite well with the actual distribution of the data. However, what one can obtain from the counterfactual considerations alone may remain too shallow for the explainee to count it as a proper explanation.

For example, the loan application system of a financial institution might return an example of such counterfactual inquiry: an applicant could obtain the loan if she was of the same age group but with a higher educational degree. This answer would be within the distribution of the data of the model and the search parameters of the query. However, the explainee cannot obtain the education desired without spending years out of one's life in order to satisfy the condition, or the applicant is beyond the age limits of a long-term loan or insurance policy. A better grasp of the conceptual situation is hence much desired, as one grows weary of having to look for extra causal rules semi-manually to expel rude turns taken in chat-bot conversations. An appeal to common sense is nowhere to be seen.

In general, the meaning of counterfactual conditionals refers to the realm of logically possible imaginary scenarios, which differ from actual ones in certain minimal respects and in which the truth-value of the antecedent of the counterfactual conditional has to be false. However, then one has to further ask with what minimal adjustments one could create a feature space (a 'possible world') in which the antecedent comes out as true while the identities of objects are preserved as a consequence of the conditional. For example,

take the sentence "If I were Chinese, I would come from the north". The first hurdle is to resolve the identity criteria for the two demonstrative pronouns expressed in the clause. The second is the qualitative measurement of the desired minimal modifications, such as their plausibility and feasibility, that are to be made to the model that has the relevant object of the first-person demonstrative seen comparably as being both non-Chinese "I" and their Chinese imaginary. However, as the systems (and the features of the surrounding worlds both in actu and in fictio) grow in complexity, finding such minimal adjustments currently defies effective long-term solutions.

3.1.3. Feature Importance and Counterfactual Examples

The process of perceiving an explanation arising from CF techniques has its beginnings in the explainee's ability to perform reasonable evaluations of possible proposed explanations and to weed out implausible ones. Given several such examples, their receiver has the latitude of choosing which of them build up a good explanation. Thus, the *interpretation* of proposed hypotheses and putting them into the perspective of the users' *goals* cannot be subtracted from the overall model design features.

Equally and unsurprisingly, the attribution of feature importance (AFI) requires the presence of such evaluative and critical reasoning components. However, to perform reasoning well means to understand some basic underlying characteristics of the system's behaviour and the fundamental dependencies of its feature maps and models. For one, if the weight for a certain feature is high, an explainee should assume that changing that feature would significantly change the model's outcome.

In other words, we get to be in a spiralling dance with the machines, however autonomous or self-supervised they are to be: arriving at any understanding of machine behaviour worth its name assumes evaluating the proposed explanations through the interpretation of their meaning in the present context and circumstance of the user, while the constraints and, thus, the nature of explanations that the machine can output depends on the past circumstances of the history of the production of the software and the underlying assumptions that reflect the values—past, present and future—that have led the development from the initial design features to the full model generation.

The difference between AFI and CF is that in AFI, the features can change according to some initially defined background values. It is the more dynamic method of the two and may for that reason be better fitted for catering explanations of the kind that respect the complexity of the interplay between images, language and reasoning from data. On the other hand, CF can provide more flexibility without hitting the ceiling of intractability as quickly as AFI.

The likeness between counterfactuals and feature importance is that they both imply the outcome of the XAI agent to be a causal attribution as an answer to a 'why' question.

What we see as a missing piece of XAI—one that sustains long-term adequacy for dynamic autonomous systems—is for the design methods of those systems to (i) effectuate self-improvement in scale, (ii) incorporate creative elements of the counterfactual meaning generation in full, and that they ultimately (iii) work with some fledgling form of commonsense reasoning and reasoning under uncertainty in which salient features of the search space are unavailable yet coherent and informative explanations are being produced. Incorporating features like (i)–(iii) into the reasoning of ML systems becomes one of the foremost future tasks. When successful, this may justify calling them the second-generation ML. When there is such genuine novelty in the explanations proposed, one might go as far as to state that the learning module is approaching the functions of reasoning that may be seen as being endowed with some capacities for abduction.

4. Abduction

Abduction was originally proposed in 1868 by Charles S. Peirce [6], and first termed 'hypothesis' to denote the type of reasoning by which creative scientific reasoning has its beginnings [13–21]. In his earlier, pre-1900 accounts, abduction was in fact a type of

inductive inference, as it was meant to sample a collection and to infer from that some general character of the collection, which is induction. Indeed, Peirce termed his earlier account 'abductive' or 'qualitative' induction. It was one among three main types of induction which he termed crude, quantitative and qualitative induction.

4.1. Attributes of Peirce's Abduction

Interestingly, Peirce made a mistake—though he corrected it later on—as many ML specialists have done, in calling this brittle type of qualitative induction by the name of abduction. Indeed, abduction proper is not induction about the characteristics but a process of forming explanatory hypotheses. It typifies the first stage of scientific inquiry, the one in which the inquirers encounter a surprising or anomalous phenomenon that creates some friction with experience no matter how weak the signal, and prompts them to look for an antecedent clause under which the phenomenon would cease to be too surprising by turning down the cognitive heat caused by the previously unknown anomaly. Peirce requires of the explanation offered by abductive inference that it must then predict the facts observed as probable outcomes of the cases in question. The process of adopting this hypothesis as an explanation is what Peirce means by abduction.

The canonical formulation of the logical form of abduction occurs in the seventh and last Harvard Lecture of 1903 [22]:

- The surprising fact, C, is observed.
- However, if A was true, C would be a matter of course.
- Hence, there is reason to suspect that A is true.

The hypothesis is the antecedent of a (supposedly) true conditional, and the conditional is the explanation of the surprising fact. Drafting an explanatory hypothesis thus amounts to finding an antecedent to complete the subjunctive conditional expressed in the second premiss. Peirce calls this form of reasoning 'retroduction' because it "starts at consequents and recedes to a conjectural antecedent from which these consequents would, or might very likely logically follow" [23]. The retroductive process of adopting the hypothesis of finding an antecedent of which the surprising fact is a consequent is the first step in inquiry [24]. The mechanics of abduction, in this sense, is that it reasons from effects to causes.

4.2. Abduction and Computation

Do our silicon friends have any propensities for such powers of abduction that the other kinds of machines, which we term the living mind or the human brain, have acquired in their evolution in nature for millions of years? Even if computers can acquire some such characteristics through complex computational feats, is it the same kind of abduction that characterises the 'creation' of hypotheses in artificial neural nets emulating synaptic connections that the hundreds of trillions of them in the human mind are presumed to perform?

One thing to note here is that abduction neither provides, nor is assumed to provide, definite or conclusive solutions, let alone best explanations to problems and observations that counter current knowledge. Rather, abduction regurgitates suggestive material for further investigation. Abductions alone do not add to present-state knowledge, the full cycle of abduction–deduction–induction does. Abduction concludes with a question and is an invitation to further inquiry. Its conclusion proposes, could X possibly be the case? Let us find out. Let us embark on a common exploration of whether X is indeed the case, given all the background knowledge, frames, scripts and interconnected evidence from related inferences that can be collated to support the proposed class of hypotheses, now accepted on probation at best.[2]

As noted, human users need to seek explanations in common engagement with machines that aim at optimising their pattern-matching tasks (and often with embarrassing results of chronic *apophenia* and pathological *overfitting* of the data). They might perform this in a fashion that we call *participatory* and *exploratory* search that characterises the desiderata of the kind of second-generation XAI that we are outlining in the present

paper. An acceptable explanation that may follow from these explorations is not only a confident assertion of what may come, let alone a referral to some non-linguistic source domain, but a result of a serious suggestion bruited by A to embark on an exploratory route concerning what may follow from the hypothesis and how further, and perhaps more decisive, evidence could be obtained to select among competitor conjectures. Maybe an ever richer and profitable conjecture is the outcome of the exploration, which may be a fine explanation. However, without a stir of the cognitive pulse by some surprising fact encountered in a problematic situation of the cognising agent, one not only learns more about the situation at hand, but may come to a trustworthy resolution to act in a certain way in all situations that are reasonably similar to each other.

A meaningful dialogue between stakeholders is thus to be established. Further requirements thus involve the emergence of the *common ground* established between human and AI agents to be engaged in purposeful communication in the first place. Typically, that would be achieved in supervised learning by the explicit input that the programmer and the maintainer provide when adjusting the weights of the nodes in the neural architecture. This can be both intentional and unintentional, and will somehow not only represent important values and choices, but also signal those values and choices to the networked system to influence its connectivity and behaviour.

In *unsupervised* learning models, the input by which the common ground is established is still there in the selection of the priors and initial conditions (training data) of the system, as well as in the subsequent software updates and hardware maintenance performed during the life cycle of the neural architectures.

In *self-supervised* learning, although much rests on the machine's abilities to work through subsequent and novel noisy data, the initial quasi-weights are assigned to the network to obtain the thing we started with. The goal of self-supervision is to prevent the convoluted network or GPT-3 from balking when faced with low-quality data, much like the idea of abduction that begins with the element of surprise and proceeds to the guess, like a toddler that learns to classify parental sounds or cats and elephants to a surprising degree of accuracy from few previous sensory exposures (and typically without having to contrast them with non-parental sounds and non-feline quadruples).

Whether (non-contrastive) self-supervision, which may be the most significant of the current ML approaches to statistical induction, can reach the levels of reasoning required of being truly originatory in the sense of strong abduction (and perhaps conversely, whether some creative elements of abduction can be identified in self-supervised learning) calls for further study that will be carried out in the sequel to the present paper. In the interim, we will suppose that human–machine interaction in all its known instances inherently is required to involve a significant amount of human-to-human communication. This is so, despite the promising fact that in efficiently self-supervised deep neural systems, processes of learning are mediated and siloed by activation patterns that emerge from deeply layered hidden structures. However, those patterns cannot configure into something that would escape the classifications and categories that the quasi-training-date has laid out in the neural layers. A possibility that remains is that an entirely new form of reasoning located somewhere between weak (statistical) and strong (creative, discovery-led) abductions might arise from elaborate iterations of habits of self-supervision; that, too, will have to be a topic of a separate discussion.

Last, we remark on the common quality that XAI frameworks may be expected to possess in relation to what abduction is intended to be. According to the common but ultimately insufficient definition of XAI, its goal is to find the best hypothesis that is to fit all available data [28]. However, this sounds too much like the question of incorporating abductive inference into ML models in terms of finding best explanations to account for the data. The identification of abduction with the inference-to-the-best-explanation(s) (IBE) is mistaken for reasons that have been explored elsewhere in formal and historical detail [29–34]. For our purposes, we take Peirce's abduction to be irreducible to induction and to Bayesian update protocols. Instead, it is the only component of any comprehensive

reasoning module in AI that is able to approach the novelty of explanations. In doing so, it is this original notion of creative abduction that paves the way for the much-needed interfaces in human–machine communication and comprehension.

4.3. XAI: Why Abduction?

How is the inference process that we can discern from XAI outputs related to abduction? The explainee needs to interpret the outputs based on the complex inner workings of the XAI agent who is to seek and find explanations that appease the elements of surprise or anomaly encountered in its performance. Again, we revisit the previous two examples, feature importance and counterfactuals, from the point of view of that the primary driver of abductions is the creative inferences.

4.3.1. Feature Importance Limits

The XAI feature importance score involves that an explainee assumes that the model's decision was made based, to some significant degree, on some specific features $(a, b, c, \ldots)$ rather than some others $(x, y, z, \ldots)$.

The explainees cannot here perform *deduction* because they do not have access to the causes of the feature weights involved in the proposed solution that could serve as the premisses of the deductive inference.

They could try *induction* by collecting all possible instances that have the same corresponding feature importance weights. In such a case, they risk computational explosion of a multi-dimensional feature space and the output would be intractable. Induction, as the canonical framework of all current expedient machine learning models, does not return a concise explanation. Making assumptions under which the decision of the machine would appear reasonable, facile and natural, given the massive data sets thus generated, would require an altogether different technique to assist a human interpreter.

4.3.2. Counterfactuals Limits

In the technique based on counterfactuals (CF), the explainee receives a limited set of examples. If the explainee is satisfied with the given set or is able to test their own input value combinations with acceptable cost, separate abductive steps might not always be needed.

When no external interactions with the input features need to be considered, the CF set can suffice as a cause-attributing explanation. For example, a simple scenario is that the user is satisfied with "You can change the values of features *a* and *b* by flipping the binary prediction outcome from 0 to 1".

At the same time, when an explainee needs to assess the relevance and sense of the input features (as in the loan example above), the complexity of transforming outputs into linguistic explanations quickly grows beyond the reach of specific binary decision models.

4.3.3. Some Connections

Given a limited certainty, making assumptions external to the knowledge of the model knowledge while observing a fraction of the data's properties and a fraction of the model behaviour have a lot to do with abductive inference—especially when the assumption is drawn according to a likelihood, relevance or suitability to the problems at hand.

For example, Bylander et al. [35] proposed, by allegedly drawing from Peirce's 1903 schema of abduction, the following formulation:

- D is a collection of data (facts, observations, givens).
- H explains D (that is, H would, if true, explain D).
- No other hypothesis can explain D as well as H does (contrastive clause).
- Therefore, H is probably true.

In case of XAI methods, AFI and CF take the role of auxiliary data collected to facilitate finding explanatory hypotheses, but are not enough to count as explanations themselves.

In the light of the above abductive schema, the XAI agent's output is a set of causal attributions, given that abduction is in its skeletal form retroductive reasoning from effects to causes.

Unfortunately, abductive conclusions are highly tentative and conjectural, and are to be accepted on probation if at all. Thus they cannot ascertain the security of the underlying reasoning processes. Although abduction is reasoning that, in connection to the human mind, can be argued to be trustworthy in certain meaningful ways and in the real-world domains [14], it does not produce the conclusions in anything like the definite or final forms of explanation that explainees would want to receive in the first place, without having to do the hard follow-up thinking by themselves.

The upshot is that causal attributions are far from *conclusive* explanations. Rather, they are initial shapes of possible explanations in the form of invitations for further investigation. Those shapes have to go through subsequent validation processes in the explainee's mind, computations, experience and experimental procedures, before attaining the status of relevant and satisfactory propositions that could count as answers to the all important 'why' questions that XAI strives to achieve.

A few remarks on the abductive schema are in order. First, while the above schema from [35] is advertised as abductive reasoning, it in fact is a case of probabilistic reasoning. For Peirce, however, probable reasoning is deductive in kind (that is, the conclusion necessarily follows with a certain degree of credence assigned to it). Abduction, in contrast, is the weakest form of inference that concludes, at best, with some *plausibility measure* of the tentative hypothesis. Fundamental abductions are not Bayesian, although abduction may enter Bayesian updates in assigning and evaluating the prior probabilities that the Bayesian equation requires to have.

Second, the second premise is a counterfactual "If H were true, it would explain D as a matter of course". While elements of what we presented as the rudiments of CF technique can be brought to play in interpreting this second premise of abduction, the problems concerning the semantics and pragmatics of counterfactual conditionals are intractable in the computational realm in scales beyond the most rudimentary and simplistic models. Additionally, here, counterfactuals typically work under probabilistic reasoning, which suggests that the interventions required are inductive and not abductive in their primary nature (see, for example, [36]).

Third, it remains to be seen whether the contrastive clause could be omitted from the premises of the abductive schema, perhaps in an analogous fashion to how contrastive self-supervised learning can be modified to perform non-contrastive self-supervised learning without much loss—or even with some improvement of the cost-effective quality of the learning results. We presume that not only can the contrastive clause be omitted in abduction, but that it would be a right thing to do in the prospected theories of automatised abduction in ML, as the question of the 'best fit' as an explanation is itself not a question of abduction alone but a question of *evidence* for how well the hypotheses proposed by abduction fare on the face of experience, such as search for counterexamples. Since the latter is the question of induction (while also mediated by deduction), finding the best fit is not to be answered by abductions alone.

5. The Common Sense Problem

The reason for the necessity of laying bare these complications on the rough path toward XAI is of course that we need to take on board the cognitive processes of an explainee and pay close attention to which parts of the given set of attributions can and which cannot (and which parts one might remain entirely agnostic to) be fully relocated to the computational part of the system.

The main problem that characterises hypothetical XAI agents is thus nothing more and nothing less than the problem of common sense. XAI agents do not really understand what they are doing, one is tempted to exclaim; they may succeed in performing some particularly well-constrained tasks but are much worse in the creation of common ground

between other agents, designers and users, necessary for the rationale of coordinated communicative intentions to be acknowledged.

In general, there is an obvious gap between a model and the levels of difficulty in capturing relevant aspects of the world and its dynamics into the agent's model. The common ground and its future scaling and leveraging could only come out of the theoretical successes and breakthroughs in how to fill in that gap.

The causal attributions can be retrieved from a model of any complexity, but the validation process (filtering) inevitably hits the ceiling of the common sense problem. Even if the model represents well the highly approximated, well-curated and efficiently mediated features of the environment, the formulation of common ground between the interacting interlocutors seems to become no easier to achieve, the reason being that the question of how to make good guesses on what the model needs to contain and what can be left out from it is one of the hardest thing to do, especially when it falls on machine intelligence to generate those guesses and to select, abstract and idealise that model of the world.

Yet for humans, common sense as an intrinsic, mutually acknowledged possession of relevant beliefs, knowledge and other second-order modalities that ground our experiences about the world is so familiar and so easily accessible (despite their obvious propensity for inconsistencies) that performing a validation of causal attributions usually goes on without consuming too much cognitive bandwidth or reflection. Common sense problem solving and inference are precisely the kinds of things that we need *not* be told about and need *not* be explicit about in the knowledge base. It is precisely the art of omitting and suppressing common sense from being explicitly encoded in the premisses that paves the way toward the genuinely novel habit formations required for efficient ampliative and abductive reasoning to shine.

So what are we asking the system designer to implement to effectuate that near-magical conversion from complex functions to efficient structures? In simpler words, what are we to ask the software engineer to do? Some of the causal attributions might be filtered with the help of knowledge that is not present in the very model that calls for explanations. Some such filterings can be defined by a set of rules, and some of these rules could be obtained from the data. Others, however, are contextual and come from norms of behaviour that do not fit well with the data-fitting paradigm of the prevailing ML archetype.

That there is some rule-based nature to the filtering of causal attributions could help eliminate out-of-distribution attributions, such as "The input feature x is too low; if it was higher, the output would be positive". Here, the feature x is correlated with the feature y, while y is also low. In this case, it would be reasonable to attribute the colligated "x and y" as the cause of a negative output, and not x alone.

There are some necessary filtering rules, too, that cannot be derived from the explained model by any activation pattern. In the case of a loan approval, for example, temporal aspects are ignored by the model. The amount of input features to such an approval mechanism is limited. A possibility is to add manually selected rules that would partially filter those attributions that, to our cerebral common-sense detection modules, are obviously ill-suited as explanations.

An even more difficult layer to track happens when the explainee has a customised profile and benefits only from a fraction of causal attributions, while all of them nevertheless are plausible, or even indistinguishable from one another in being equally plausible.

In such cases, the explanation from an agent could be admitted by an explainee once it is *actionable*, that is, it gives grounds for acting in certain ways in certain kinds of circumstances. From the point of view of Peirce's conception of abduction, actionable explanations are precisely those that invite further thinking and active inquiry in order to find more about the real status of that explanation. It comes close to acquiring a *belief* on how things might stand, a belief strong enough to serve as a proxy for justified action.

There may ultimately be a need for abductions that can infer a final or unique explanation out of a set of causal attributions proposed and filtered by an XAI agent. However, as we struggle to cope with the complexity of the world and its representation filtered by

an explainee's individual, social, cultural and ethical contexts and values in order even to begin to address the problem of *final*, *ultimate* or *conclusive* explanations, we are forced to leave these hefty matters for now and only nurse them with the following thought.

The steps towards better abductive performance of deep learning could be effected by novel filtering algorithms. The applicability of those techniques is limited and currently confined to what, in the literature, is termed *selective* abduction. Mere filtering does not address well the side of the *generative* (creative) abductive reasoning, one that produces genuinely new conclusions to serve as antecedents of the counterfactual premise to explain away the peculiarities of observed matters and surprising facts.[3] The programme of generative (creative) abduction and real discovery, in contrast, would at once have to address the issue of narrowing down the space of hypotheses that takes place not merely as a mechanism of filtering, but as a mechanism of only paying attention and expending significant computational and cognitive resources on determining explanations that would be accepted as being plausible and feasible, and having some practical bearings by the explainee in the first place. We feel that this task requires a modal system of higher-order reasoning that is able to produce clauses that encode the dictates of some universal common sense.[4] That side of the story—with an allusion to [42]—remains to be written in almost its entirety in the theory and practice of fundamental data sciences.

6. Conclusions

It is not hard to picture what is at stake in these issues. Detecting unusual features in satellite imagery or medical scans becomes much more valuable if the flags are at the same time interspersed with reasons as to why the system 'thinks' that a recognised anomaly or suspected pathology should indeed be handled as such in the further proceedings. Explaining from footage that some behaviour of a person in a crowd is dangerously anti-normative is hard, and not least because such behaviour is likely to be novel and unforeseen and not present in the world's interconnected data repositories on past human behaviour. Averting collateral disasters hinges on future machines' abductive abilities to generate novel and creative explanations that can look beyond the available data, can sample the future scenarios by smart guesses, and come up with narratives of how one might get from here to there or else prevent dreadful scenarios from ever materialising in the first place. None of this is yet within the realm of reliable machine performance.

Hence, the successful generation of satisfactory, feasible and at best conclusive explanations, is a touchstone of the truly novel abilities expected of future and emerging AI systems. Are machines trustworthy companions in reasoning from effects to causes, that can be endowed with the tasks of proposing hypotheses on probation, hypotheses that are novel but plausible and hence worthy of further investment of time, energy and money to see if they are a matter of course?

Breaking the code of the logic of genuinely generative abductive reasoning—one that invents reliable, trustworthy, well-aligned and non-confabulatory *post-hoc* reasons and explanations of observed and anomalous phenomena as the conclusion of that mode of inference—thus becomes of the upmost importance in the future agenda of ML and its parental data sciences. Current XAI methods, which aim at providing causal attributions as their outputs to satisfy end-user's requests for explanation, are inadequate for the systems to reach these wider aims—not to mention being rather perilous ones too, when fully autonomous system behaviour is an issue.

Even so, we have argued that the problem of the validation of a set of attributions is not well translatable to the model that represents the agent. The agent model is supposed to make abductive turns in its overall inferential module, but that requires interventions external to the knowledge-bases of the models that are to be explained. With counterfactual explanations, for one, it may be possible to define filtering rules that become incrementally closer to the goal states of the knowledge model, but in general, the final choice of the real cause(s) and, respectively, of the validation of an explanation, rests on the end user's

expertise, resources and powers of reasoning, even in the much fan-fared self-supervised cases of late.

Without such proper interaction between the explainee (the inquirer) and the explainer (the ML system that strives to represent chunks of nature), the stated goals of XAI cannot be asymptotically approached. In this paper, we have sketched what a participatory, exploratory, second-generation XAI ought to look like when more attention gets to be drawn at its reasoning modules by which satisfactory explanations are to be extracted from data plus their creative manipulation and abstraction. We argued that interdisciplinary XAI research exemplifies core elements of scientific and everyday inquiries that look for satisfactory answers to 'why' questions. This will not really happen without abduction. Thus, an improved understanding of autonomous machine behaviour mandates going beyond the inductive paradigm to the science of the dialectics between abductive and inductive stages of reasoning mediated by deduction. To reach that higher goal of AI inquiry, in turn, calls for undoing some of the fundaments of data sciences that typify much of current ML research and its attitude that tends to perceive such an end as little more than what an instrumentally rational, inductive enterprise can pursue.

Author Contributions: Conceptualisation, K.M. and A.-V.P.; methodology, K.M. and A.-V.P.; formal analysis, K.M. and A.-V.P.; investigation, K.M. and A.-V.P.; resources, K.M. and A.-V.P.; writing—original draft preparation, K.M. and A.-V.P.; writing—review and editing, K.M. and A.-V.P. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Estonian Research Council's Personal Research Grant PUT 1305 ("Abduction in the Age of Fundamental Uncertainty", PI A.-V. Pietarinen); Chinese National Funding of Social Sciences "The Historical Evolution of Logical Vocabulary and Research on Philosophical Issues" (Grant no. 20& ZD046).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the Philosophy of Science & Technology PhD class of 2021 at TalTech for lively discussions on the draft of the present paper.

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

Notes

1 A reviewer wisely reminds us of Hintikka's admonition of years ago, namely that explanation remains one of the most theoretically intractable of our philosophical concepts: it is common enough for a human explainer to do a perfectly sound job of explaining himself without, even so, enlarging the explainee's understanding. Is universal intelligibility a condition on sound explanation making? The objection might be this. Suppose the explainee wants to know why the explainer voted as he did, and receives in answer the reply, "Because of their better understanding of monetary policy". It could be true that that decision was arrived at by the causal forces to which human cognition is subject, but to press him to recapitulate those impacts in his answer invites a confusion of questions. It is one thing to ask someone why he voted as he did, but quite another to ask him to specify the causal workings of what brought him to that action. It might be that the causal forces that regulate the respondent's preferences, together with such knowledge of monetary things which his belief-producing mechanisms may have endowed him with, caused the vote in question. However, since the human agent is largely innocent of the causal engineering that brings such things to pass, it is simply unavailing to demand that he reveal it when a 'why' question is presented to him. Yes, most of the time we call upon our System 1 to produce responses as purported explanations, as confabulatory and self-deceptive though they are, as well as being noisy and extremely fallible in their claim for knowledge of relevant causal factors. Ignorant of the antedating powers that caused those utterances to arise from the undifferentiated soup of beliefs, motives, desires and learned behaviours, an answer produced to the 'why' question is to be assessed with such a universal human condition in mind.

2 On the interrogative construal of Peirce's abduction and its peculiar underling pragmatic and modal logics, see [25–27].

3 On the proposed distinction between selective and generative abduction, see, for example, [37–39].

4 Again, a hint from Peirce may be his practical, diagrammatic approach to modalities; see [40] and [41].

References

1. Mooney, R.J. Integrating Abduction and Induction in Machine Learning. In *Abduction and Induction*; Flach, P.A., Kakas, A.C., Eds.; Applied Logic Series 18; Springer: Dordrecht, The Netherlands, 2000. [CrossRef]
2. Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, 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]
3. von Wright, G.H. *Explanation and Understanding*; Cornell University Press: Cornell, NY, USA, 1971.
4. Hintikka, J. *Semantics of Questions and Questions of Semantics*; Acta Philosophical Fennica: Helsinki, Finland, 1975.
5. Hintikka, J. *Knowledge and Belief: An Introduction to Logic of the Two Notions*; Cornell University Press: Ithaca, NY, USA, 1962.
6. Peirce, C.S. On the Natural Classification of Arguments. *Proc. Am. Acad. Arts Sci.* **1868**, *7*, 261–287. [CrossRef]
7. Recanati, F. Unarticulated Constituents. *Linguist. Philos.* **2002**, *25*, 299–345. [CrossRef]
8. Barocas, S.; Selbst, A.D.; Raghavan, M. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 80–89.
9. Stepin, I.; Alonso, J.M.; Catala, A.; Pereira-Fariña, M. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* **2021**, *9*, 11974–12001. [CrossRef]
10. Lewis, D. *Counterfactuals*; Blackwell Publishing: Oxford, UK, 1973.
11. Sanford, D.H. *If P Then Q: Conditionals and the Foundations of Reasoning*; Routledge: London, UK, 1989.
12. Stalnaker, R.C. A Theory of Conditionals. In *Studies in Logical Theory*; Rescher, N., Ed.; Blackwell: Oxford, UK, 1968; pp. 98–112.
13. Bellucci, F.; Pietarinen, A.-V. The Iconic Moment: Towards a Peircean theory of scientific imagination and abductive reasoning. In *Epistemology, Knowledge, and the Impact of Interaction*; Pombo, O., Nepomuceno, A., Redmond, J., Eds.; Springer: Dordrecht, The Netherlands, 2016; pp. 463–481.
14. Bellucci, F.; Pietarinen, A.-V. Peirce on the Justification of Abduction. *Stud. Hist. Philos. Sci. Part A* **2020**, *84*, 12–19. [CrossRef] [PubMed]
15. Bellucci, F.; Pietarinen, A.-V. Methodeutic of Abduction. In *Abduction in Cognition and Action*; Shook, J., Paavola, S., Eds.; Springer: Cham, Switzerland, 2021; pp. 107–127.
16. Fann, K.T. *Peirce's Theory of Abduction*; Nijhoff: The Hague, The Netherlands, 1970.
17. Gabbay, D.M.; Woods, J. *The Reach of Abduction. Insight and Trial*; Elsevier: Amsterdam, The Netherlands, 2005.
18. Gabbay, D.M.; Woods, J. Advice on Abductive Logic. *Log. J. IGPL* **2006**, *14*, 189–219. [CrossRef]
19. Hintikka, J. What Is Abduction? The Fundamental Problem of Contemporary Epistemology. *Trans. Charles S. Peirce Soc.* **1998**, *34*, 503–534.
20. Hintikka, J. *Socratic Epistemology: Knowledge: Explorations of Knowledge-Seeking through Questioning*; Cambridge University Press: Cambridge, MA, USA, 2007.
21. Kapitan, T. Peirce and the Structure of Abductive Inference. In *Studies in the Logic of Charles Peirce*; Houser, N., Roberts, D.D., van Evra, J., Eds.; Indiana University Press: Bloomington, IN, USA, 1997; pp. 477–496.
22. Peirce, C.S. *The Collected Papers of Charles S. Peirce*; 8 vols.; Hartshorne, C., Weiss, P., Burks, A.W., Eds.; Harvard University Press: Cambridge, MA, USA, 1931–1966.
23. Peirce, C.S. Volume 1: History and Applications, In *Logic of the Future: Writings on Existential Graphs*; Pietarinen, A.-V., Ed.; Mouton De Gruyter: Berlin, Germany; Boston, MA, USA, 2019.
24. Pietarinen, A.-V.; Bellucci, F. New Light on Peirce's Conceptions of Retroduction, Deduction, and Scientific Reasoning. *Int. Stud. Philos. Sci.* **2014**, *28*, 353–373. [CrossRef]
25. Chiffi, D.; Pietarinen, A.-V. Abductive Inference within a Pragmatic Framework. *Synthese* **2020**, *197*, 2507–2523. [CrossRef]
26. Ma, M.; Pietarinen, A.-V. A Dynamic Approach to Peirce's Interrogative Construal of Abductive Logic. *IFCoLog J. Log. Appl.* **2015**, *3*, 73–104.
27. Ma, M.; Pietarinen, A.-V. Let Us Investigate! Dynamic Conjecture-Making as the Formal Logic of Abduction. *J. Philos. Log.* **2018**, *47*, 913–945. [CrossRef]
28. Hoffman, R.; Klein, G. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intell. Syst.* **2017**, *32*, 68–73. [CrossRef]
29. Campos, D. On the Distinction between Peirce's Abduction and Lipton's Inference to the Best Explanation. *Synthese* **2011**, *180*, 419–442. [CrossRef]
30. Mcauliffe, W.H.B. How Did Abduction Get Confused with Inference to the Best Explanation? *Trans. Charles S. Peirce Soc.* **2015**, *51*, 300–319. [CrossRef]
31. Pietarinen, A.-V. The Science to Save Us from Philosophy of Science. *Axiomathes* **2014**, *25*, 149–166. [CrossRef]
32. Woods, J. Recent Developments in Abductive Logic. *Stud. Hist. Philos. Sci. Part A* **2011**, *42*, 240–244. [CrossRef]
33. Woods, J. Cognitive Economics and the Logic of Abduction. *Rev. Symb. Log.* **2012**, *5*, 148–161. [CrossRef]
34. Woods, J. Reorienting the Logic of Abduction. In *Springer Handbook of Model-Based Reasoning*; Magnani, L., Bertolotti, T., Eds.; Springer: Berlin, Germany, 2017; pp. 137–150.
35. Bylander, T.; Allemang, D.; Tanner, M.; Josephson, J. The Computational Complexity of Abduction. *Artif. Intell.* **1991**, *49*, 25–60. [CrossRef]

36. Crupi, R.; Castelnovo, A.; Regoli, D.; Gonzalez, B.S.M. Counterfactual Explanations as Interventions in Latent Space. *arXiv* **2021**, arXiv:2106.07754.
37. Magnani, L. *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*; Springer: Dordrecht, The Netherlands, 2009.
38. Magnani, L. *The Abductive Structure of Scientific Creativity*; Springer: Cham, Switzerland, 2017.
39. Park, W. *Abduction in Context: The Conjectural Dynamics of Scientific Reasoning*; Springer: Dordrecht, The Netherlands, 2017.
40. Ma, M.; Pietarinen, A.-V. Gamma Graph Calculi for Modal Logics. *Synthese* **2017**, *195*, 3621. [CrossRef]
41. Peirce, C.S. Volume 3/1: Pragmaticism, In *Logic of the Future: Writings on Existential Graphs*; Pietarinen, A.-V., Ed.; Mouton De Gruyter: Berlin, Germany; Boston, MA, USA, 2022.
42. Larson, E.J. *The Myth of Artificial Intelligence Why Computers Can't Think the Way We Do*; Harvard University Press: Cambridge, MA, USA, 2021.