Editorial:

From Explainable Artificial Intelligence (xAI) to Understandable Artificial Intelligence (uAI)

I. Introduction

In this editorial, we argue that the Artificial Intelligence (AI) community needs to escape the “trap” of Explainable Artificial Intelligence (xAI) by growing more research on Understandable Artificial Intelligence (uAI). We provocatively term xAI a trap because it has caused some AI researchers to see it as the “end” rather than a “means”. We will discuss why uAI is a better way forward and present a framework for uAI to define research directions that go beyond xAI. Let us first share where this concept emerged before introducing it.

On the 1st of July 2024, Prof. Abbass presented an online Invited Talk at the IEEE World Congress on Computational Intelligence (IEEE WCCI 2024) on “Explaining Explainable Artificial Intelligence”. This was followed by a two-hour panel to discuss Explainable Artificial Intelligence (xAI) on the 3rd of July. The panellists were the authors of this editorial, Chaired by Prof. Garibaldi. The first half of the panel was a debate between Prof. Gegov and Prof. Abbass on xAI, where Prof. Gegov was given the “affirmative” role for arguing that xAI is a necessary condition for trust, safety, responsibility and accountability. In contrast, Prof. Abbass was given the “negative” role, in which he argued instead that xAI is not enough and the community needs to expand the scope of research to the broader concept of uAI. An engaging Q&A with the audience and other panellists followed the debate. The second half of the panel included presentations from Prof. Sousa on an industry perspective on xAI, Prof. Crockett on human-centred xAI and legislation, and Prof. Kaymak on algorithmic explanation. In the remainder of this editorial, we will present a refined account of the points made during the invited talk and the panel.

II. Explainable Artificial Intelligence

The need for explanation has persisted in the AI field since its inception. Before deep learning, the concept of an explanation was integral to the symbolic AI school in areas such as expert systems and knowledge-based decision support systems. In the connectionism school, rule extraction from neural networks saw significant growth in the nineties, with growing literature exploring different approaches to explain neural networks. Such approaches suggested transformations for neural network architectures to convert them into decision rules, tables, and trees. In the computational intelligence community, explanation has been central in methodologies such as learning classifier systems and genetic programming and, indeed, at the heart of the whole field of fuzzy systems.

The rise of deep learning almost caused a shift in the perception of AI from being a field in academic labs to a commodity ready for end-users to create wealth. This revolution caused a spike in the need for users to understand AI, including AI models, develop confidence in their output, and trust in adopting them, especially in essential problems. DARPA sensed the signal as early as 2016 and established a program in Explainable AI (xAI).

Gunning and Aha [1] summarise DARPA’s journey. They emphasised the significance of xAI in their statement that “explaining AI will be essential if users are to understand, appropriately trust, and effectively manage these artificially intelligent partners.” ([1, p.44]) They continued to state how DARPA defines xAI “as AI systems that can explain their rationale to a human user, characterise their strengths and weaknesses, and convey an understanding of how they will behave in the future.” ([1, p.44]) Gunning and Aha categorised DARPA’s program using three questions: “(1) how to produce more explainable models, (2) how to design explanation interfaces, and (3) how to understand the psychological requirements for effective explanations." ([1, p.s45]) They then discussed the three strategies for developing explainable models: (1) deep explanation, (2) interpretable models, primarily causal models, and (3) model induction. It is worth noting that the most frequent word in Gunning and Aha's paper was "understand" (and its derivations). The emphasis in this paragraph is ours and will be revisited in our discussion below.

DARPA's promotion of xAI contributed to an exponential growth in the number of papers on the topic. We categorise this literature into three categories. The first includes conceptual and literature review papers. The second contains papers that attempt to design algorithms to provide explanations, especially for black-box models such as deep learning. The third includes papers on machine reasoning, planning, and (multi-)agent systems, where studies use symbolic representation, natural language processing, and visualisations as organic representations for explanations.

The above xAI literature has been compounded by confusion about terminologies. For example, it became common to see researchers confusing or interchanging the terms "interpretability" and "explainability." It is not uncommon to see many algorithmic papers considering the provision of an explanation as the end goal; that is, the problem is solved if the algorithm can explain its output without necessarily considering the type of user such an explanation is appropriate for, let alone how one explanation interacts with subsequent ones in the formation of user mental models.

The premise of producing explanations and then integrating these explanations to offer understanding was successful in creating an extensive literature on explanation, but unfortunately, little advancement was made on "understanding". Perhaps this approach fails because explanations cannot be decoupled from understanding; the two are tightly coupled and must be addressed together. For example, the separation between the production of explanations and psychological factors impacting human understanding leads to suboptimal integration, where explanations become solutions waiting for problems. These papers assume that the problem is solved by designing an algorithm that provides some extra information called explanations; they do not define the "actual" problem! In other words, DARPA's promotion of xAI was motivated by the three aims mentioned by Gunning and Aha: enabling human understanding, improving human trust, and managing partners in team arrangements. Unfortunately, these aims somehow got buried in the volume of literature that focused on explanations as an end.

The literature often argues that explainability is essential in building trust amongst non-technical end-users. Yet, technical explanations generated through XAI cannot be easily understood by people who are not fully aware of the use of AI in their daily lives. Let us consider, for example, the needs of a wider pool of stakeholders, citizens such as marginalised communities and those suffering from digital poverty. These groups are less likely to be interested in technical explanations. Instead, they might be more interested in a plausible understanding of why an automated decision was made using their data and how such a decision affects them personally. Clear, compelling, jargon-free communication of a decision appropriate for the educational level of users' stakeholder pool would enhance understanding and promote future engagement and participation. Despite these somewhat cynical comments, xAI has successfully advanced explanations of black-box models such as deep learning, and the credit goes to DARPA for creating the xAI movement. Admittedly, significant challenges remain on this algorithmic level. For example, generating explanations for reinforcement learning algorithms, time series data, and dynamical systems is non-trivial. Nevertheless, we should not lose sight of the primary motivation; the aim of xAI is not explainability but understandability.

III. Understandable Artificial Intelligence

"To understand" is a multi-faceted concept. Researchers in philosophy and epistemology have been debating this concept for centuries. Nevertheless, we will provide concrete definitions. Our aim is not to provide a universally acceptable definition of understanding but to offer clarity using a technological lens on an unambiguous function definition that could be described concisely to a system engineer to implement and assess it.

Christoph Baumberger [2] described three understanding types: propositional, interrogative and objectual. Propositional understanding concerns facts; for example, "We understand that the backpropagation algorithm could get stuck in local optima". Anyone who has studied artificial neural networks may be able to understand this assertion.

Interrogative understanding is explanatory; for example, "we understand that the backpropagation algorithm gets stuck in a local optimum due to its reliance on first-degree gradient information with a deterministic hill-climbing approach". Here, we understand the reason. While interrogative understanding is the closest form of understanding to explanation, Baumberger [2] convincingly argues that knowing the reasons does not automatically entail understanding the reasons and vice-versa. In other words, an explanation does not necessarily entail understanding, while understanding the reasons could occur without being "told" the reason. Take the reason why backpropagation gets stuck in a local optimum. Lack of a background in calculus may mean that a human who now (after being provided with an explanation) knows the reason but still does not understand the reason. Similarly, a student who uses visualisation to inspect the backpropagation algorithm as it traverses over a loss function may understand why it gets stuck in a local optimum without being told the reason.

Objectual understanding is understanding the subject; for example, "we understand artificial neural networks". The premise here is that one develops objectual understanding through a body of knowledge. Once again, objectual understanding could be achieved without necessarily being presented with explanations in the form of reasons or arguments; it could be achieved, for example, by introducing a set of facts on neural networks, a video showing a neural network in action, or a book providing the mathematical basis for neural networks.

The above epistemological summary shows that xAI is neither necessary nor sufficient for uAI. The mere existence of an explanation does not automatically lead to understanding, which could be developed through forms other than explanations. But what is understanding, after all? Is it achievable? Is it measurable? If so, what are the tenants of uAI?

IV. An Information Lens for Understanding

Philosophy, epistemology and hermeneutics literature present diverse views on understanding. A common thread exists that understanding is not the mere availability of information but requires the agent to grasp, digest, and be mentally comfortable with this information. In other words, understanding assumes that the agent is self-aware and conscious and can self-judge the additional knowledge it receives.

One can then map out the process of understanding four pieces of technology. Figure 1 summarises the four components of understanding technology. In the remainder of this section, we will unfold each of them.

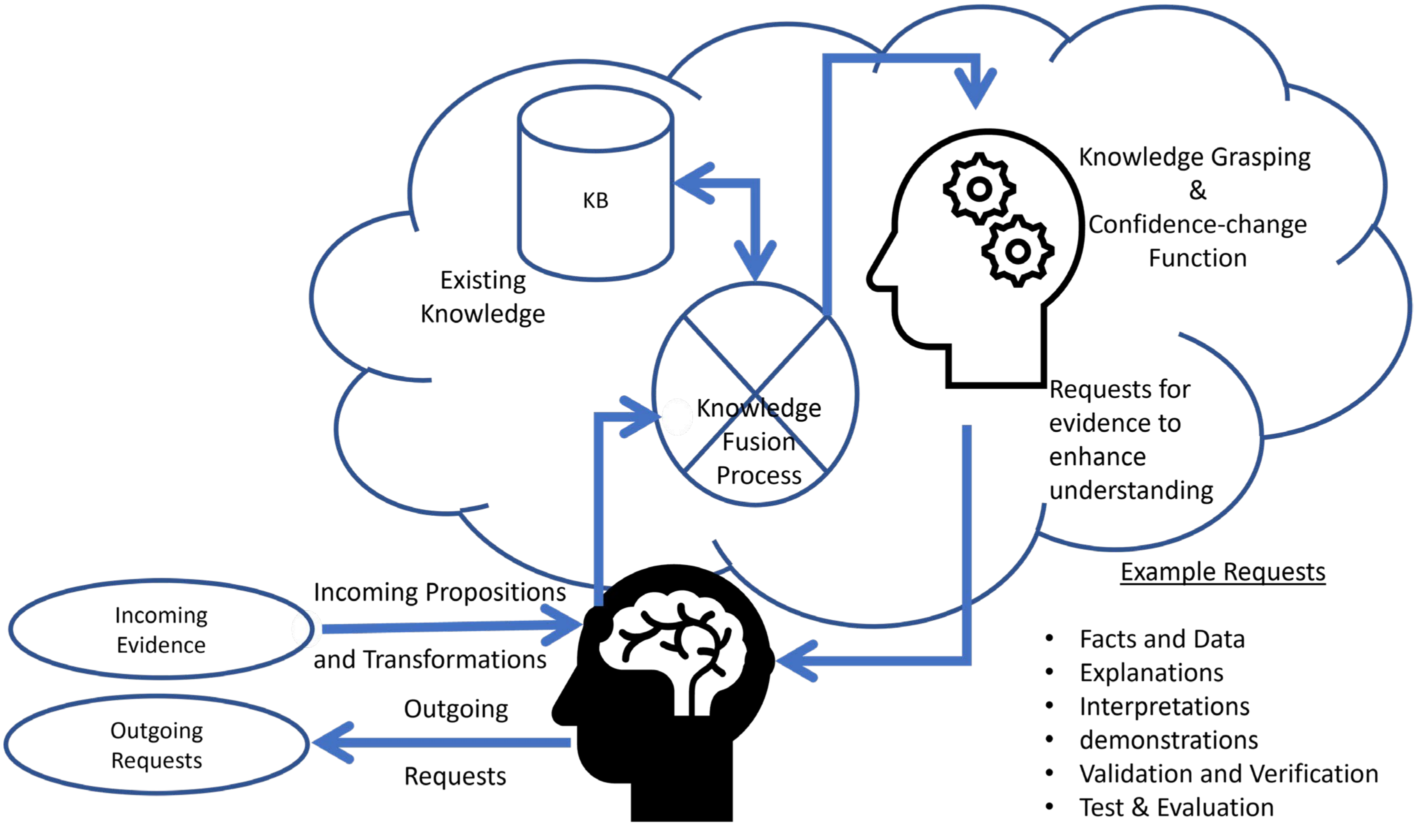

Fig. 1. Understanding Process.

A. Incoming Propositions and Transformations

The first component is a technology or a medium responsible for the arrival of evidence, which can take many modalities, including processing modalities (messages, data, information and knowledge) and representational modalities (e.g. text, audio, video). Without evidence reaching the agent, understanding can not take place. It is essential to clarify that this necessary condition is about “evidence”, not “explanations”. An agent may receive evidence as an explanation; however, it may also receive evidence where the agent needs to interpret, reorder, and process the evidence before extracting explanations. Some pieces of evidence are purely data-oriented, which can be used to validate parts of an agent’s mental model.

Incoming messages to the agent need to be interpreted into the two categories of knowledge that the agent has: propositions and transformations. Both categories could be represented symbolically (i.e. symbolic logic) or numerically (i.e. neural networks). An agent may receive an assertion about the world or self through its sensors or communication channels. These assertions could be in the form of propositions or transformations. The agent may use a consistency judgement function (a similarity metric such as distance or scoring functions) to assess if the incoming message is consistent with its existing knowledge. If the judgement suggests that the incoming message sufficiently deviates from, or is inconsistent with, the internal expertise or if the incoming message falls into a sub-space where the agent has little to no knowledge, the agent will use a response judgement function to assess whether it can and should seek evidence to support the incoming message and to weight it against what it knows already, or to discard it. Supporting evidence could be requested and provided in various forms, including causal explanations, demonstrations, experimental evidence, historical evidence, etc.

B. Existing Knowledge

The second piece of technology resides within the agent in the form of the agent's knowledge base. Understanding can not occur if the agent lacks background knowledge, prior experience, internal beliefs, and encoded knowledge to grasp incoming information. Understanding assumes and demands that the agent has some existing knowledge in place. This knowledge set can generally be categorised into knowledge about the world, self, and knowledge about knowledge processing. One may see the third category as meta-knowledge, but we will flatten the hierarchy to avoid a vacuous cycle of knowledge about knowledge about knowledge, and so forth. While technologically, there could be an efficiency dividend gleaned from organising knowledge into hierarchies, flattening the knowledge base offers more flexibility in representation, where, by default, knowledge is seen to operate on knowledge; thus, knowledge unifies the subject and object. Such a view extends upstream to meta-knowledge and downstream to treating pieces of information and data as pieces of knowledge. Thus, the world for an agent consists of propositions (facts, assertions, hypotheses, beliefs) and transformations, including deductive/inductive/abductive inference models, represented in various forms like rules, neural networks, probability graphs, and so on.

C. Knowledge Fusion Process

The third technology is a fusion process, where incoming knowledge is integrated with prior knowledge. This knowledge fusion technology enables the agent to mix, diffuse, compress and optimise its knowledge base. Whether the agent needs to assess the consistency of an incoming message with its internal knowledge or the agent has accepted the incoming message, it needs to integrate it with its expertise and apply a knowledge fusion process.

The consistency of an incoming message with an agent's existing knowledge could take two forms. The first is a direct similarity between the incoming knowledge and existing knowledge. The second is an inference using existing knowledge to derive the nearest state to the incoming knowledge. In this second case, knowledge fusion will need to be applied to derive the nearest state to the incoming knowledge.

Fusion could be seen as transitioning from two or more states to one state. The input states are what the agent knows already and what the agent is receiving; every piece of incoming knowledge is an instantiation of a feature in the knowledge state space. The fused knowledge is equally a new state in the knowledge state space. This framework extends naturally to more complex internal representations, such as counterfactual reasoning, in which the agent considers alternate states of knowledge, including contradictions and dissimilarities.

V. Confidence-change Function

The first three components mentioned in the previous sections can be mechanical. Research is technologically advanced in all three components; we have the algorithms and technologies for AI systems to communicate knowledge, store knowledge, and update and fuse knowledge. However, does this mean AI can understand?

Knowledge fusion falls short of understanding. One of the main challenges is the need for self-awareness and self-assessment regarding the compatibility of updated knowledge with old knowledge. A sub-challenge is the lack of functional definitions of what "digesting knowledge" means for a human. When we grasp knowledge, is it merely that we have a logical chain to infer what we receive from what we know (i.e. all we need is an inference mechanism), or does it extend to judgement due to benchmarking fused knowledge against meta-knowledge, values and deep beliefs? This grasping process is less understood; therefore, perhaps unsurprisingly, little literature exists on the functional implementation of "grasping", let alone the premise that understanding implies a level of mental comfort after fusing incoming knowledge.

The above discussion suggests an agent needs a confidence change function; however, while this function is necessary alone, it is insufficient. Understanding needs internal confirmation within an agent to decide that it is comfortable with its fused knowledge; that is, it can assess if its degree of confidence in the change that has occurred is sufficiently high to be satisfied. This could be a complex technological piece, including ensuring the integrity of updated knowledge, connectivity (each inference of something the agent knows is consistent with what it knows), functions to assess the change in an agent's degree of confidence, and a few. We will only emphasise here the change in the degree of confidence in each piece of knowledge that has been affected by the fusion process before and after knowledge fusion.

The confidence-change function has two roles: to accurately estimate the impact of incoming messages on an agent's existing knowledge and to assess the agent's readiness level to digest the incoming messages. A simple example should demonstrate both uses.

Consider an agent holding the proposition that the sun rises from the Northeast with confidence 0.95, along with transformations with high prediction accuracy. Assume this agent's experience has been based on winter data from the Southern hemisphere and summer data from the Northern hemisphere. The agent then receives a message that the sun rises from the southeast from a world traveller agent experiencing the summer in the Southern hemisphere. The confidence-change function will flag this as a piece of knowledge inconsistent with internal knowledge; the distance is high. The agent may request data points from the sender, who then delivers ten thousand observations covering the Southern and Northern hemispheres in both seasons. The agent trains new transformations on this data and then compares its previous transformation predictions with the predictions of the latest transformations. The agent notices that all predictions match winter data from the Southern hemisphere and summer data from the Northern hemisphere; the incoming evidence/data and the resultant transformations are consistent with what it knows from before but differ in sampling new sub-spaces. The confidence-change function will return a pair of values, representing the value of the additional knowledge and the confidence level in its consistency with existing knowledge.

In the example above, the value of the additional knowledge and confidence level will be high because the incoming message covers sub-spaces that the agent has not seen before and does not contradict what the agent already knows.

VI. Understandability and Usability

Using an information lens, we argue that the above four building blocks form the minimum viable technologically feasible model of understanding in AI agents. Moreover, these building blocks could equally be applied to humans to inspect where understanding is easy or hard and the cases for such states. Considering these four learning building blocks, uAI is concerned with ensuring that the AI is ready to provide the information needed for understanding, that is, the necessary information to support each building block in performing efficiently.

One can see the "u" in uAI to both denote "understandable" and "usable". For an AI to be usable, it needs to be understandable and making it understandable improves its utility and usability. uAI must be designed with the four building blocks mentioned above in mind. When uAI makes a prediction or recommends an action, it needs to consider the knowledge base of the receiver agent or a model of it. When multiple outputs are available, uAI needs to favour the one closest to what the receiving end knows. When the output is distant from the other agent's knowledge, uAI needs to provide the evidence required by the receiving end to understand next to the production. As shown in Figure 1, evidence could take many forms, including causal explanations, arguments, a subset of the positive and negative data, scenarios, the model used for generating the output, or an interpretation.

The receiving agent may have a limited form of representation. For example, a human decision-maker may prefer a simple visual representation of the decision or an argument in plain English. uAI needs to provide at least the required data to produce these representations. It is even better when uAI provides these pieces of evidence in a receiver-ready form. However, providing the evidence necessary for an agent to understand is insufficient. At least two extra conditions are required: continuous feedback to ensure that the provided proof aligns with the agent's needs and constant assessment of understanding to assure understanding; the latter is where the fourth building block plays an important role.

A more sophisticated uAI will need to provide evidence in a form to facilitate the knowledge fusion process and confidence- change function. For example, the explaining agent may need to provide the winter data from the Southern hemisphere to demonstrate to the receiving agent that it learnt the same thing. When the explainer does not consider the receiver, these data would be seen as redundant and unnecessary. Instead, these data would reinforce that part of the experience is shared, which could promote trust among the agents.

Our proposed four building blocks also constrain the nature of assertions and knowledge updates. For example, the existence of a consistency judgment function assumes that the similarity of an assertion to existing knowledge can be assessed. Propositions that do not satisfy this property can thus be not allowed. Overall, this implies that explanatory propositions must be commensurable in some way and that it must be possible to challenge an explanation with alternative propositions. Similarly, uAI research must establish which are admissible in the knowledge fusion and update steps using the confidence change function. Indeed, this reflects that uAI research cannot be considered independent of epistemological and philosophical underpinnings.

VII. Conclusion

This editorial shares with the community a perspective by the authors that was presented at the 2024 IEEE WCCI conference. We have argued that “understanding” was the original motivation for xAI. However, the research field has grown to see explanations as the end rather than the means to understanding. We argue that significant research opportunities exist for uAI, in which (1) explanation is only one type of evidence needed for understanding and other forms of evidence, such as statistical evidence and demonstrations, need research, and (2) algorithms for explanation need to be designed with “understanding” in mind. We have suggested four building blocks for the minimum viable solution to develop understanding within an artificial agent and suggested that these four building blocks need to guide the formation of explanations by AI.

References

[1] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[2] C. Baumberger, “Types of understanding: Their nature and their relation to knowledge,” *Conceptus*, vol. 40, no. 98, pp. 67–88, 2014.