FELSEFE DÜNYASI

2024 YAZ/SUMMER Sayı/Issue: 79

FELSEFE / DÜŞÜNCE DERGİSİ

Yerel, süreli ve hakemli bir dergidir.

ISSN 1301-0875

Sahibi/Publisher

Türk Felsefe Derneği Adına Başkan
Prof. Dr. Murtaza Korlaelçi

Türk Felsefe Derneği mensubu tüm Öğretim üyeleri (Prof. Dr., Doç. Dr., Dr. Öğr. Üyesi) Felsefe Dünyası'nın Danışma Kurulu/Hakem Heyetinin doğal üyesidir.

Felsefe Dünyası, her yıl Temmuz ve Aralık aylarında yayımlanır. 2004 yılından itibaren Philosopher's Index ve TÜBİTAK ULAKBIM/TR DİZİN tarafından dizinlenmektedir.

Felsefe Dünyası is a refereed journal and is published biannually. It is indexed by Philosopher's Index and TUBITAK ULAKBIM/TR DİZİN since 2004.

Editör/Editor

Prof. Dr. Hasan Yücel Başdemir (Ankara Üniversitesi)

Yazı Kurulu/Editorial Board

Prof. Dr. Murtaza Korlaelçi (Ankara Üniversitesi)
Prof. Dr. Sema Önal (Kırıkkale Üniversitesi)
Doç. Dr. Fatih Özkan (Ankara Hacı Bayram Veli Üniversitesi)
Doç. Dr. Muhammet Enes Kala (Ankara Yıldırım Beyazıt Üniversitesi)
Dr. Öğr. Üyesi Aynur Tunç (Ankara Yıldırım Beyazıt Üniversitesi)
Arş. Gör. Ahmet Hamdi İşcan (Ankara Üniversitesi)

Alan Editörleri/Section Editors

Prof. Dr. Ahmet Emre Dağtaşoğlu (Trakya Üniversitesi)
Doç. Dr. Fatih Özkan (Ankara Hacı Bayram Veli Üniversitesi)
Doç. Dr. Mehmet Ata Az (Ankara Yıldırım Beyazıt Üniversitesi)
Doç. Dr. Sebile Başok Diş (Necmettin Erbakan Üniversitesi)
Doç. Dr. Nihat Durmaz (Selçuk Üniversitesi)
Dr. Mehtap Doğan (Ankara Yıldırım Beyazıt Üniversitesi)
Dr. Muhammet Çelik (Ankara Sosyal Bilimler Üniversitesi)
Dr. Kenan Tekin (Boğaziçi Üniversitesi)
Dr. Nazan Yeşilkaya (Şırnak Üniversitesi)

Yazım ve Dil Editörleri/Spelling and Language Editors

Zehra Eroğlu (Ankara Üniversitesi)
Abdussamet Şimşek (Ankara Sosyal Bilimler Üniversitesi)
Hatice İpek Keskin (Ankara Sosyal Bilimler Üniversitesi)

Fiyatı/Price: 300,00 TL | **Basım Tarihi :** Temmuz 2024, 300 Adet

Adres/Adress

Necatibey Caddesi No: 8/122 Çankaya/ANKARA
Tel: 0 (312) 231 54 40
https://dergipark.org.tr/tr/pub/felsefedunyasi

Hesap No / Account No: Vakıf Bank Kızılay Şubesi
IBAN: TR82 0001 5001 5800 7288 3364 51

Tasarım / Design: Turku Ajans

Baskı / Printed: Uzun Dijital
Zübeyde Hanım, İstanbul Çarşısı, İstanbul Cd. No:48 D:48, 06070 Altındağ/Ankara
Tel: (0312) 341 36 67 | **Sertifika No:** 47865

Derginin online versiyonu ücretsizdir.
The online version of the journal is free of charge.

THE MACHINE AS AN AUTONOMOUS EXPLANATORY AGENT

OTONOM BİR AÇIKLAYICI FAİL OLARAK MAKİNE

Dilek YARGAN
Dr., University of Rostock Institute of Philosophy Department of Philosophy,
ORCID: 0000-0001-9618-6740, e-mail: dilek.yargan@uni-rostock.de

Felsefe Dünyası Dergisi, Sayı: 79, 2024, ss. 265-279.

Geliş Tarihi: 21.05.2024 | Kabul Tarihi: 12.07.2024
DOI: 10.58634/felsefedunyasi.1487376
Araştırma Makalesi - Research Article

Abstract

The holy grail of Artificial Intelligence (AI) is to transform the machine into an agent that can decide, make inferences, cluster the contents, predict, recommend, and exhibit similar higher cognitive faculties. The prowess of Large Language Models (LLMs) serves as evidence: they enable seamless natural language communication and widespread use across various fields by swiftly processing unstructured data and handling diverse datasets with agility. However, in order to be competent in the fields of science and industry, an agent with such capabilities must be reliable, i.e., accountable for its decisions and actions, which is a per se attribute of an autonomous agent. In this respect, this paper aims to determine whether state-of-the-art technologies have already created an autonomous explanatory agent or are paving the way for the machine to become an autonomous explanatory agent. To achieve this, the paper is structured as follows: The first part investigates the types and levels of explanations in explanation models, providing a foundation for understanding the nature of explanations in everyday life. The second part explores explanations in the context of artificial intelligence, focusing on types of explanatory systems in the research field of eXplainable AI (XAI). The third part delves into whether and to what extent the state-of-the-art machine learning models function as autonomous explanatory agents, based on the exploration in the second part and considering the field of Human-Computer Interaction.

Keywords: machine explanation, autonomous agent, XAI, ontology, HAI

Öz

Yapay zekâ çalışmalarının nihai amacı, makineyi karar verebilen, çıkarım yapabilen, öngörebilen, tavsiyelerde bulunabilen ve diğer yüksek bilişsel işlevleri gerçekleştirebilen otonom bir faile/eyleyiciye dönüştürmektir. Büyük Dil Modellerinin sergilediği üstün yetenekler, bu amacın gerçekleştiğini ya da gerçekleşmesine ramak kaldığının adeta bir kanıtıdır, zira yapısal olmayan verileri işleme hızları ve veri çeşitliliğini çevikçe işleyebilme yetenekleri çeşitli alanlardaki geniş kullanımlarını mümkün kıldığından makine-insan arasındaki doğal dil iletişimini kesintisiz hâle getirmiştir. Ancak, bilim ve endüstride yetkin olabilmek için, bu tür yeteneklere sahip olan bir eyleyicinin güvenilir, yani eylemlerini ve aldığı kararları açıklayabilir olması gereklidir, ki hesap verebilme otonom bir failin başat niteliğidir. Bu bağlamda, bu makale, mevcut teknolojilerin halihazırda otonom bir açıklayıcı fail oluşturup oluşturmadıklarını veya makinenin otonom bir açıklayıcı fail olmasına zemin hazırlayıp hazırlamadıklarını belirlemeyi amaçlamaktadır. Çalışmanın ilk bölümü açıklama modellerindeki açıklama türlerini ve seviyelerini araştırarak günlük yaşamdaki açıklamaların doğasını anlamak için bir temel ortaya koyar. İkinci bölüm, Açıklanabilir Yapay Zekâ alanındaki açıklayıcı sistem türlerine odaklanarak yapay zekâ çalışmalarındaki açıklama modellerini inceler. Çalışmanın devamında ise, ikinci bölümdeki incelemeye ve İnsan-Bilgisayar Etkileşimi alanına dayanarak, güncel makine öğrenmesi modellerinin otonom açıklayıcı fail işlevi olup olmadığını ve ne ölçüde olduğunu araştırılır.

Anahtar Kelimeler: makine açıklamaları, otonom fail, XAI, ontoloji, IBE, IFE

Yazarlar dergide yayınlanan çalışmalarının telif hakkına sahiptirler ve çalışmaları CC BY-NC 4.0 lisansı altında açık erişim olarak yayımlanmaktadır.
The authors own the copyright of their work published in the journal and their work is published as open access under the CC BY-NC 4.0 licence.

Introduction

The prowess of large language models (LLMs) represents a state-of-the-art advancement in deep learning, becoming an integral part of everyday life in recent years. Whether or not they will evolve into artificial general intelligence, it is evident that these models can communicate in natural language with ease due to their enhanced semantic power. This power is bolstered by semantic tools and resources, such as ontologies, knowledge graphs, and thesauri, in addition to other statistical and mathematical technologies, like hyper-heuristics. Accordingly, the use of these tools in industry and science has become widespread.

In industry, the present-day technologies of the Internet, data analyses, Big Data, robotics, and alike introduce true and full automation from production to distribution processes. The assemblage of these technologies is believed to transform automated factories into autonomous factories. This transformation means increasing profits, decreasing costs, improving customer experience, maintaining newly introduced raw materials, optimizing lifetime value, and other market issues are held by the intensive assistance of autonomous machines. This is Industry 4.0: establishing smart object networking and autonomous process management, where the interplay between the physical and digital realms becomes a vital new dimension of manufacturing and production processes. (GTAI, n.d.). Such intelligence paves the way for autonomous decision-making in all aspects of marketing, namely in the design, production, operation, and service of products.

Since the advent of computers and their involvement in scientific knowledge production, the machine has become an indispensable tool that has changed the way of doing science: without the software, specific data could never have been collected, analyzed, or used to draw conclusions. Scientists have to rely on the results that the machine generates; therefore, the machine has become an indispensable component of scientific knowledge production. Current improvements in the technology of experimentation and measurement yield a vast amount of scientific data. What has been revolutionary in science is that the machine is involved in data generation and analyzing processes; in other words, the machine has become indispensable in scientific knowledge production. Furthermore, the impact of Big Data, data analytics, and integration of LLMs in science, namely the upcoming scientific revolution, is when the machine behaves like a `scientist:' it can systematically observe, conduct experiments, do the reasoning upon its findings, construct hypotheses, and test hypotheses. In other words, the machine becomes autonomous.

Accordingly, the most significant impact of state-of-the-art technologies is the transformation of the machine into an *agent*—a system actively seeking to fulfill a collection of goals in a complex and dynamic environment (Maes, 1993). Such an agent can participate in industrial workflow and scientific knowledge production. Therefore, these technologies aim to create algorithms that can decide, make inferences, cluster the contents, predict, recommend, and exhibit similar higher cognitive faculties to achieve the goals at hand. Moreover, in these fields, such an agent is called *autonomous*, an active agent that independently determines how to relate given data to background data to achieve its goals successfully. Ultimately, these technologies make the machine understand Big Data; namely, for a given goal/situation, the agent should decide itself by structuring, analyzing, and interpreting a given context and then making inferences on the final product while producing particular models for the context.

That said, however, the alleged autonomous agents need to be accountable to be effective. Without explanations, we can never be sure that the machine's discoveries, decisions, and other cognitive-like operations are true, correct, and reliable. This paper aims to determine whether state-of-the-art technologies pave the way for the machine to become an explanatory agent by investigating three questions. First, we will understand what an explanation is (section 2); second, we will explore this notion in the context of artificial intelligence (section 3); and lastly, we will answer the question of what it means to be an autonomous explanatory agent (section 4). We will then extend our conclusion to relate to Human-Agent Interaction.

What is an Explanation?

In everyday usage, the term 'explanation' refers to making something known or explicit. More systematically, however, explanation refers to making something known or explicit via models that can vary in their structures. For instance, in the deductive-nomological model, a scientific explanation is a sound deductive argument that follows from a particular class of true premises, at least one of which is a law of nature; without that, the derivation is invalid. Alternatively, consider statistical models where co-variations are used to construct explanations. Some models also include social aspects of explanation. Hilton (1990) describes explanation as a conversational model, where Grice's maxims of explanations (quality, quantity, relation, and manner) rule over a conversation.

Along with the structure of explanation, the types and levels of explanation should be considered in explanation models. Types of explanation refer

to the categorization of explanations of phenomena. For instance, Aristotle's *Four Causes* model provides explanations from four distinct classes: material, efficient, formal, and final causes; Dennett suggests three stances for explaining objects: physical, design, and intention; Marr proposes three points from which a computational problem can be understood: computational, representational, and hardware levels; and, Kass and Leake group explanations of anomalies into three types: intentional, material, and social (Miller, 2019: 10).

Furthermore, levels of explanation refer to providing explanations according to the domain-knowledge level of the explainee. As the level of knowledge and experience increases, the level of abstraction of the explanations or the number of technical terms used in the examples will differ. For instance, a medical doctor examining an X-ray explains her findings to a colleague differently than an untrained patient. Therefore, the explanation should be tailored to the explainee's background knowledge, which the explainer should consider.

Levels of explanation are also related to the process of cognition of the explainer and the explainee (Gunning and Aha, 2019). Miller (2019) surveys three types of cognitive processes used in the explanations. The first is *causal connection*, which involves the explainer generating the causes of the phenomenon by collecting data and manipulating it with prior knowledge and observations. Investigating several studies, Miller shows that explainers choose different causal chains for explaining the same phenomenon. Social norms, gaining experience, explanatory types, and identifying counterfactual cases during explaining can affect identifying the sequence of critical causal connections.

The second one is *explanation selection*, which involves the explainer selecting a subset of the identified causes to provide *the* explanation. Related to pragmatic goals, selecting such a subset is connected to the cognitive biases of the explainers. Miller provides plenty of examples illustrating how people employ different criteria for choosing important, relevant causes. For instance, identifying necessary and sufficient conditions are to be used as explanatory causes; alternatively, explanation criteria can be reduced to functional and mechanical explanations.

The third type of cognitive process used in explanation is explanation evaluation, which is how the explainee assesses the explanations they receive. People use different criteria to determine whether the provided explanation is reasonable. For instance, Grice's maxims of explanation evaluation (quality, quantity, relation, and manner) help the explainee judge the expla-

nations. In addition, coherence, simplicity, and generality are among the most commonly used criteria (Miller, 2019).

Consequently, there are various criteria to consider before constructing an explanatory model for science and everyday usage. When an explanation contains humans as explainers or explainees, their mental models, prior beliefs, experiences, expertise, and different degrees of understanding about the context affect the course of explanatory dialogue. It may be worthwhile to consider whether formalizing such models for the machine is feasible. In the following lines, we will investigate what explanation is in artificial intelligence and which models are used for this purpose.

What Is Explanation in Artificial Intelligence?

Explanation in artificially intelligent systems, often referred to as 'eXplainable Artificial Intelligence' (XAI), is a field of research that aims at producing explainable models, with effective explanation techniques, that provide human-understandable explanations to intelligent system's decisions, recommendations, predictions, or actions (Gunning and Aha, 2019). The demand for studies in XAI has increased dramatically in the last few years because machine learning (ML) models, particularly LLMs, are increasingly being employed to make predictions in critical fields, such as security, finance, and health. This is because most used ML models, i.e., deep learning models, are notoriously black-box systems; namely, the end-users –humans– cannot understand how the model produces its results. Therefore, there must be explanations supporting the results of these models so that, for humans, the system's decisions, recommendations, predictions, or actions are transparent, justifiable, and legitimate. To illustrate this crucial point, consider a scenario where a person is killed by a drone strike during a riot. At the very least, the intelligent system *must* explain and justify the reasons behind its decision and subsequent action to *target* and *kill* that particular person.

A Remark on Explainability and Interoperability

A distinction between 'explanation' and 'interoperability' is necessary for further discussions. Arrieta et al. (2020) stress the misuse of the terms 'interpretability' and 'explainability' in the literature. They define interpretability as "a passive characteristic of a model referring to the level at which a given model makes sense for a human observer" and explainability as "an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions" (Arrieta et al., 2020: 4-5). In this context, interpretability can also be expressed as

transparency. Hall et al. (2019) differentiate explanation and interpretation as different tasks. Gunning and Aha (2019) note that DARPA's XAI Project is deliberately named "explainable" rather than interpretable, comprehensible, or transparent AI. They argue that the term "explainable" reflects DARPA's aim for building more human-understandable AI systems. On the other hand, Biran and Cotton (2017) and Miller (2019) equalize interpretability with explainability. In this work, we focus on the definitions in the explanatory AI models provided by the researchers.

Types of Exploratory Systems

Although all researchers agree that an XAI model is designed for humans to understand how an intelligent system reaches a conclusion, disparate explanation needs and different conceptual connotations lead to various meta-interpretations of XAI.

Biran and Cotton (2017) survey explainability in the ML literature around two key aspects: interpretability and justification. An intelligent system is interpretable if a human can understand its operations through either introspection or a self-generated explanation. Whereas, a justification defends why a decision is a good one without stating how the decision is generated. Justifications are crucial for non-interpretable systems. In line with this and based on trend analysis in the ML literature, the authors categorize XAI research into two main branches: (a) interpretable models and (b) prediction interpretation and justification. Interpretable models are readily interpretable by humans or are inherently interpretable. For instance, decision trees and association rules can be explained through reasoning by humans, or Bayesian models can be interpreted when studied in detail. The weights of the attributes or the probability of the results are observable so that they are interpretable.

In prediction interpretation and justification, on the other hand, the predictions of complex models, viz. non-interpretable models, are interpreted. Indeed, these models produce justifications for the prediction. For instance, support vector machine classifiers can justify neural network models, which are notoriously non-interpretable, by extracting conjunctive rules operating on a small subset of features (Biran and Cotton, 2017: 3)

This survey was extended by a thorough analysis of Arrieta et al. (2020). At the beginning of the analysis, they use a more widely accepted classification of XAI: transparent models and post-hoc explainability. The transparent models are also called interpretable models and are assumed transparent if it is understandable by itself. Logistic regression, decision trees, k-nearest

neighbor, and rule-based learning are some examples of transparent models. The post-hoc explainability techniques are developed when a model cannot provide any insight to a human. In other words, a particular XAI technique is devised to explain the decisions of an already-developed model. That is why these techniques are also called post-modeling explainability. There are text explanations, visualizations, local explanations, explanations by example, feature relevance estimation, explanations by simplification, and feature relevance, to name some post-hoc explainability techniques.

Doran, Schulz, and Besold (2017) present four concepts of XAI concerning types of explanations: opaque systems, interpretable systems, comprehensible systems, and explainable systems. Opaque systems are genuine black boxes where no mappings from inputs to outputs are visible to the user. These systems are like oracles that generate some results without providing any rationale or sequence of thought. The user cannot get an explanation from the algorithmic mechanisms of opaque systems. Interpretable systems, on the other hand, exhibit their algorithmic mechanisms to a degree. A user can mathematically investigate how inputs are mapped to outputs. Nonetheless, the user needs to know the technical details of the mapping in advance. For instance, a regression model or a support vector machine provides an equation or a set of equations whose coefficients can be compared with each other to understand the mappings. Next, comprehensible systems produce symbols as by-products that allow the user to relate the properties of the inputs to their outputs. For instance, visualizations of the predictions assist the users in evaluating the intelligent model. Similar to the interpretable systems, the user must compile and comprehend the symbols in these systems. The user's expertise level affects the interpretation of the relations between inputs, symbols, and outputs. That is to say, different users may comprehend different things from the symbols and the models. High-dimensional data visualizations and receptive field visualization on convolutional neural networks are examples of comprehensible models.

In the last two systems, so-called explanations are comprehended and interpreted by the user. However, these explanations require human post-processing, where humans, who may be either model developers or domain expert users, serve as experts. These experts evaluate crucial XAI traits such as confidence, trust, safety, ethicality, and fairness through their reasoning abilities in symbol comprehension and mathematical interpretations.

Lastly, Doran et al. (2017: 7) introduce truly explainable systems as the ultimate notion of XAI, which can formulate "a line of reasoning that ex-

plains the decision-making process of a model using human-understandable features of the input data." Unlike the last two systems, explainable systems yield explanations autonomously without requiring collaboration with human analysts who may introduce errors and different explanations during the explanation generation process.

To address such issues and eliminate human-generated explanations, an (truly) explainable system features a reasoning engine that combines symbols emitted by a comprehensible system with a domain knowledge base, representing the relationships between symbols. Thus, functioning as automated reasoning machines, these models act as explanatory agents.

The notion of (truly) explainable systems by Doran et al. (2017) reminds the efforts of the DARPA XAI Program. Researchers from various universities, companies, and institutes work on creating ML-based explainable models that "enable end-users to understand, appropriately trust, and effectively manage the emerging generation of AI systems" (Gunning and Aha, 2019: 45). To this end, the researchers (i) produce as many as explainable models with new or modified ML techniques, (ii) design effective explanation interfaces that work on the explanation models. Since the interface is aimed to be designed with principles and techniques of Human-Computer Interaction (HCI), the DARPA XAI model is planned to behave as an explanatory agent. That is to say, the researchers also focus on (iii) understanding the psychological requirements for convincing explanations to make the model more human-understandable. As such, the DARPA XAI Program can be regarded as an explainable system. Nevertheless, this is not the case, as we will see shortly.

Hitherto, we have seen that there are three types of XAI. The first one is interpretable, which covers the interpretable systems of Doran et al. (2017) and the transparent models, where straightforward what-if scenarios can be conveyed in a decision tree by a domain expert. The second type is explainable, which covers the comprehensible systems of Doran et al. (2017) and post hoc explainability. The DARPA XAI Program fits in here since the researchers are developing an XAI system that inherits various explanatory models built with ML techniques. The last one, what Doran et al. (2017) call (truly) explainable systems, inherits reasoning built with knowledgebases, which can provide semantic features in the explanations.

The defenders of semantic tools, including ontologies, knowledgebases, and taxonomies, in XAI models, such as Baclawski et al. (2020), utilize the hierarchy of explanation types proposed by Doran et al. (2017) to illustrate an example. Suppose that one applies for a loan, which is then denied. This per-

son asks the responsible-intelligent agent, "Why was my application for the loan denied?" No matter a human or a machine, the agent must express the explanation in natural language. Consider the following examples for each type. The interpretable type might respond, "According to the logistic regression method we use for making decisions, your application result is 23%." However, it is impossible for an inexpert to understand what 23% signifies. The explainable type, on the other hand, would reply, "The system denies loan applicants with low bank account balances." Although the previous answer was a sort of identification of a statistical result, this answer provides a sort of explanation. The explainee may further want to know what low bank account balances mean. The agent of an explainable system cannot clarify the issue further since their knowledge is limited to explaining how inputs are mapped to outputs. Lastly, the reasoning type would provide an answer like, "According to the documents you provided, your expenses are more than your income, so the system does not approve loans to those who cannot show evidence of being able to pay them off." Baclawski et al. (2020: 92) state that

> The first type ... fails to relate the decision to the context in which the decision was made. The second type ... [expresses the decision] in terms of the customer context ... Yet the second type only explains the function without explaining why that function is being used. The third type ... explains why that function was used by the bank. The third type uses formal reasoning to infer the rationale from the other types of explanation ... What is not shown in the example for the Reasoning type is that it should allow for an interaction with the customer with the goal of achieving customer acceptance of the explanation.

Hence, an explainer can address any follow-up questions of the explainee to clarify within a particular context. This indicates twofold importance. Firstly, a context is open to any question types -what, why, who, how, and alike. An explainer should provide answers to all these types of questions. As such, the explanatory models are not limited to causal explanations: non-causal models are as crucial as causal ones. Secondly, an explanatory model enabling follow-up questions aims at a cooperative conversation between the explainer and the explainee, such as questions like: Why not go to Heidelberg? What if I choose X? Why should I invest in gold? Who is responsible for this mistake? Should I consider withdrawing my money? In essence, explanatory dialogue occurs between two intelligent agents. However, the counterparts have different goals: the explainer aims to provide a proper answer and generate trust, whereas the explainee wants to understand the given decision, recommendation, or action. Although explanations

are contextual, such interaction not only needs to reveal the facts but also the foils, the counterfactual cases, which are cases that can occur instead of the already happened case (Miller, 2019: 3). The explainer may encounter hypothetical cases. Yet, the foils must have meaningful relations with the context. For instance, a patient can ask an intelligent system, "Why do you recommend me taking those pills rather than having surgery?"

What is an Autonomous Explanatory Agent?

In this section, we will list the characteristics of an autonomous explanatory agent in light of the explanatory models we have examined so far.

To further our inquiry, it is essential to recognize that the machine, functioning as an intelligent agent, provides explanations to humans, who also act as intelligent agents. Such rationalizations must be explained by the explainer rather than requiring an explainable explainer; in other words, we need a system that explains the results, not an explainable system. In this respect, an explaining system should be able to question and answer in natural language and whose expressions should be easily understood by the lay user. Structure and levels of explanation play a critical role in designing an intelligent system capable of facilitating explanatory dialogues.

An intelligent system should be designed to function as an explanatory agent (Baclawski et al., 2020). Therefore, explainability must be an integral part of its design. Such a design should also result in an autonomous explanatory agent capable of independently determining how to relate a given question to background data for a successful explanation. Autonomy, in this context, consists in the ability to explain questions within a complex, dynamic environment. The machine should provide explanations with respect to the explanees' background, motivations, expectations, and capacities. It also should facilitate a back-and-forth conversation, adapting the data gathered throughout the exchange.

Understanding language or *capturing meaning* is essential for tracing back the reasons behind the results provided by the models (Bender and Koller, 2020). The key point is that explainability relies on semantically reliable architecture. Related to the previous paragraph, therefore, an autonomous explanatory agent must be designed to manipulate data semantically in a dynamic environment.

To sum up, an autonomous explanatory agent is not merely an explainable system but one that actively engages in meaningful, adaptive, and context-aware dialogues with human agents. By integrating characteristics such as interactive communication and semantic understanding, such an agent can tailor its explanations to the unique needs of each user. Its adapt-

ability and ability to handle dynamic environments ensure that the agent remains relevant, effective, and reliable as conditions change. In the following, we will conclude our investigation against these characteristics.

Conclusion

This work investigates whether state-of-the-art technologies can be regarded as autonomous explanatory agents. The investigation basically focuses on three aspects: (i) as agents, (ii) as explainers, and (iii) as autonomous explanatory agents. For the sake of argument and without delving deeply into the notion of an agent, these models qualify as agents because they interact with human users and fulfill various goals assigned by these users in complex and dynamic environments.

The machine, as the explainer, provides explanations to humans, the explainees. As discussed in Section 3.2, algorithmic transparency, post-hoc explainability, or interpretable models cannot serve as explanatory agents. An explaining system should be able to engage in question-and-answer interactions in natural language, with its expressions easily understood by users from various backgrounds and interests; hence, the structure and levels of explanation play a critical role in designing an intelligent system capable of facilitating explanatory dialogues. That said, on the one hand, given prompt instructions, an LLM can adjust the structure and level of an explanation. However, these instructions are always static, so such a model cannot serve as a true explainer. And above all, due to its design capabilities, an LLM is not an explaining system *per se*. Yet, some argue that an LLM is able to change and improve its explanation capacity over time, which is one of the behaviors of an active agent (Cf. Maes, 1993). Furthermore, it can facilitate a back-and-forth conversation, adapting based on the data gathered throughout the interactions. For instance, MemGPT (Hou et al., 2024) and Think-in-Memory (TiM) (Liu et al., 2023) mimic human-like long-term memory, designed to enhance the recall abilities of LLMs and provide long-term human-machine interactions.

Given that the XAI models we have covered so far are intended to be integrated into intelligent systems as modules, we end up with a model that generates answers, an integrated module that acts as long-term memory, and another module that explains the overall generated answer. The question, then, is whether the combination of these models constitutes an autonomous explanatory agent. A concise response is in the negative: Whether modular or as a whole, an ML model cannot meet explanation requirements due to its statistical nature. For instance, statistical algorithms

cluster concepts instead of classifying them; consequently, even sophisticated deep learning models struggle to represent and generalize the intricate structure of the world reliably. Moreover, while applications of LLMs, such as in scientific domains, strive to provide explanations that transcend the contexts present and those of the explainee, instances of hallucinations serve as evidence that LLMs cannot consistently deliver accurate explanations despite their capacity to exceed the context provided in pretraining and prompts. Therefore, explainability must be an integral part of its design. Such a design should result in an autonomous explanatory agent capable of independently determining how to relate a given question to background data for a successful explanation. The agent is expected to have the ability to explain questions within a complex, dynamic environment; the machine should provide explanations with respect to the explainees' background, mental models, expertise, motivations, expectations, and capacities, if possible; and to automate explanations within a specific context by generating a coherent and meaningful train of reasoning.

Accomplishing this necessitates the machine's ability to process data semantically. However, fulfilling this requirement is challenging with existing methods, as structured data reliant on semantic tools can be limiting, while semantically deficient machine learning models may lack the sophistication needed for effective and reliable explanation generation. Thus, state-of-the-art technologies cannot stand as autonomous explanatory agents by themselves, not only because of the complexity of building an explanation model but rather due to their lack of a semantically reliable architecture. To overcome this problem, semantic tools, such as ontologies and knowledge graphs, have already been introduced as part of the design of XAI models (Cf. Kommineni, König-Ries, and Samuel, 2024). Ontologies, in particular, provide a common framework for modeling explainable systems that interact with humans on everyday tasks (Baclawski et al., 2020). Integrating ontologies into LLMs plays a key role in autonomous explanatory systems by representing and reasoning about the world. Indeed, ontologies are crucial for improving interoperability between systems by creating a common framework for representing interpretations and explanations. Indeed, the Retrieval-Augmented Generation (RAG) architecture is frequently employed to integrate these tools into LLMs. However, neither the powerful but static nature of neural networks nor the dynamic but labor-intensive limitations of semantic networks can be avoided.

It is noteworthy that the design of an autonomous explanatory agent falls within the research domain of Human-Agent Interaction (HAI), a field

often confused with Human-Computer Interaction (HCI). Miller (2019) underscores that human-generated explanations are influenced by biases and social expectations, emphasizing the necessity to address these factors for enhanced human-machine interactions within XAI. Understanding the cognitive processes involved in explanation, particularly causal connections and explanation selection, is crucial for shaping the design of the machine as an autonomous explanatory agent. As articulated by Miller (2019: 2), HAI represents the intersection of artificial intelligence, social science, and HCI.

Historically, HAI has been a pivotal pursuit in the field. In his seminal work "Man-Computer Symbiosis," Licklider (1960) outlined key prerequisites for establishing an effective collaborative relationship between humans and machine, including advancements in memory organization and user-friendly programming languages. These enhancements are essential for fostering a symbiotic partnership wherein humans and machine can seamlessly collaborate in problem-solving and decision-making processes. Given current trends, HAI research should prioritize the development of semantic-operating architectures capable of processing unstructured data to address the challenges inherent in HAI advancement.

There is another final note worth mentioning, although a thorough discussion on it is beyond the scope and aim of this paper. In light of the discussion above, three key implications arise. First, epistemologically, these models challenge traditional notions of knowledge and understanding. The categorical differences between a human and a model, which are autonomous explanatory agents, necessitate a new framework to force us to reconsider what it means to "know" and "understand" and how to justify the explanations based on such knowledge. Second, these two entities should be ontologically reconsidered with care from the point of the notion of agency, which brings along considerations regarding the definition of living. Lastly, ethically, deploying such an agent raises issues of responsibility and accountability. Determining the locus of responsibility for decisions made by an autonomous agent is complex. Ensuring these models respect human cognitive and contextual capacities is essential, highlighting concerns about fairness, bias, and potential manipulation. To sum up, an autonomous explanatory agent compels us to reconsider fundamental philosophical concepts surrounding knowledge, understanding, agency, and moral responsibility.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.

Baclawski, K., Bennett, M., Berg-Cross, G., Fritzsche, D., Sharma, R., Singer, J., ... & Whitten, D. (2020). Ontology Summit 2019 communiqué: Explanations. *Applied Ontology*, 15(1), 91-107.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185-5198.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8, No. 1, 8-13.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint* arXiv:1710.00794.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), 44-58.

GTAI. (n.d.). *Industrie 4.0,* Retrieved February 28, 2024, from https://www.gtai.de/en/invest/industries/industrial-production/industrie-4-0.

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., & Preece, A. (2019). A systematic method to understand requirements for explainable AI (XAI) systems. In *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019),* Macau, China (Vol. 11).

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1), 65.

Hou, Y., Tamoto, H., & Miyashita, H. (2024). "My agent understands me better": Integrating Dynamic Human-like Memory Recall and Consolidation in LLM-Based Agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1-7.

Kommineni, V. K., König-Ries, B., & Samuel, S. (2024). From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. *arXiv preprint* arXiv:2403.08345.

Licklider, J. C. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1), 4-11.

Liu, L., Yang, X., Shen, Y., Hu, B., Zhang, Z., Gu, J., & Zhang, G. (2023). Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*.

Maes, P. (1993). Modeling Adaptive Autonomous Agents. *Artificial Life Journal*, 1(1-2), 135-162.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.