Minds and Machines
https://doi.org/10.1007/s11023-022-09603-z

Explainable Artificial Intelligence in Data Science

From Foundational Issues Towards Socio-technical Considerations

Joaquín Borrego-Díaz[1] · **Juan Galán-Páez**[1]

Received: 25 October 2021 / Accepted: 17 April 2022

Abstract
A widespread need to explain the behavior and outcomes of AI-based systems has emerged, due to their ubiquitous presence. Thus, providing renewed momentum to the relatively new research area of eXplainable AI (XAI). Nowadays, the importance of XAI lies in the fact that the increasing control transference to this kind of system for decision making -or, at least, its use for assisting executive stakeholders- already affects many sensitive realms (as in Politics, Social Sciences, or Law). The decision-making power handover to opaque AI systems makes mandatory explaining those, primarily in application scenarios where the stakeholders are unaware of both the high technology applied and the basic principles governing the technological solutions. The issue should not be reduced to a merely technical problem; the explainer would be compelled to transmit richer knowledge about the system (including its role within the informational ecosystem where he/she works). To achieve such an aim, the explainer could exploit, if necessary, practices from other scientific and humanistic areas. The first aim of the paper is to emphasize and justify the need for a multidisciplinary approach that is beneficiated from part of the scientific and philosophical corpus on Explaining, underscoring the particular nuances of the issue within the field of Data Science. The second objective is to develop some arguments justifying the authors' bet by a more relevant role of ideas inspired by, on the one hand, formal techniques from Knowledge Representation and Reasoning, and on the other hand, the modeling of human reasoning when facing the explanation. This way, explaining modeling practices would seek a sound balance between the pure technical justification and the explainer-explainee agreement.

Keywords Explainable Artificial Intelligence · Data science · Complex systems · Bounded rationality · Symbolic Artificial Intelligence

✉ Joaquín Borrego-Díaz
jborrego@us.es

Extended author information available on the last page of the article

Published online: 12 May 2022

Some problems are so complex that you have to be highly intelligent and well informed just to be undecided about them.
Laurence J. Peter.

1 Introduction

«1» Generally speaking, Artificial Intelligence (AI) plays two roles in Decision-Making. The first one is as an assistant to the process itself, by providing information through *inference* (e.g., a profile about a subject or situation) to the (human) agent responsible for the decision. The second one is as agents with actual autonomy, both in decision-making and the execution itself (e.g. deleting videos with unauthorized use of copyrighted music). In any case, AI technology is not exploited as an isolated artifact, it is imbricated in a broader treatment of information, often socio-technical systems that include humans (AI engineers, data scientists, scientific experts, stakeholders, etc.). Therefore, monitoring of the overall system where such modules are embedded becomes a critical concern. It will need some specifications of the system behavior, notoriously in several Data Science (DS) environments such as Big Data (BD), the Internet of Things, and Cloud Computing. Sometimes it is sufficient to get an explanation of the particular outcome or decision, what would be called a *local explaining*.

«2» The need is not global, although it is mandatory in scenarios where people's rights are affected. Stakeholders should ask for some form of certification, traceability and evaluation of applicability and performance (van de Poel, 2020). Likewise, AI-based systems are used for executive decision-making advice on particular issues in Complex Systems (CS) that seem harmless and may not be so. Cases as varied as social networks, social dynamics, prediction of urban dynamics, etc. may not be critical, and still, they could impact user's rights.

«3» AI systems are paradigmatic technical artifacts, *objects with a technical function and a physical structure consciously designed, produced, and used by humans to realize such a function* (Kroes et al., 2006). But nor every artifact that makes an automated decision is an AI system, nor every AI is Machine Learning (ML), and nor everything announced as AI is, in fact, AI-based. There is an evident hype on the subject; the term is often used within marketing strategies to justify business decisions. However, the use of AI-based modules for empowering another kind of system makes, indeed, that the latter inherit the potential explainability needs from the former. Such inheritance could happen regardless of the three main complexity levels of research in Explainable AI (XAI) (Doran et al., 2017): comprehensible systems that emit symbols enabling user-driven explanations of how a conclusions are reached, interpretable systems where users can mathematically analyze its algorithmic mechanisms, and opaque systems that offer no insight into its algorithmic mechanisms. The latter is the most troublemaking one. It could be stated that, in the context of XAI, such systems should be understood as *High Technology*, according to D. Ihde's definition (Ihde, 2010, p. 58):

> Complex and intertwined systems that while are understood through scientifically derived theories, their components are esoteric (nor do we understand

their function) although we know that they are the result of complex and scientifically determined processes and finally we concentrate some information on tolerance and internal organization.

As it is frequent in Complex Artificial Intelligent Systems (CAIS), the (formalized principles from) scientific models are hidden inside the software. Thus the high level of understanding required becomes difficult to acquire (as Ihde's notion points out).

«4» Rather than limiting itself to the details of the technology, in some cases, the explanation should also show the outcomes that the AI-based solution can bring. Also, how would its impact on the business be as a whole? Such requirements turn the explanation into a product of an interactive scenario, a socio-technical system framed by the interaction between the data scientist -assisted by a CAIS- and the stakeholders. This is a product of the current socio-technological ecosystem where AI systems are increasingly used to support decision-making of institutions and governments. It is not a completely new setting in AI (e.g. the well-established Knowledge Engineering), only is becoming much more relevant due to the new systems and players involved. For example, it sought solutions that must satisfy the need for transparency in decision making that affects the citizenry (de Fine Licht & de Fine Licht, 2020). In fact, it is a challenge how to achieve the general public perceiving AI-based decision-making as a producer of legitimate and acceptable solutions. The process transparency could augment such a perception, damaged by the technology evolution. The ubiquitous deployment of increasingly complex AI systems challenges humans' confidence in their performance, experimental correctness, and validity. In addition, there is the fact that *Trust* in High Technology comes in a different belief to the AI engineer than to a politician (as an extreme example). Broaching this issue by using tools transferred and adapted from Knowledge Engineering may be insufficient.

«5» As argued by Miller (2019), the creation of explainable intelligent systems requires addressing some issues. Firstly, those that come from the consideration of Explaining as the product of the interactivity between humans and the (automated or semi-automated) AI system. Secondly, the design of representations supporting the articulation of the explanations is required. To these requirements others should be added -somewhat distant from IA- that would affect other dimensions, such as the social (inter-agent). Weld and Bansal (2018) required for a *good* explanation to be simple, easy to understand, faithful (accurate), and conveying the true cause of the event. They shape the balancing problem between two demands: is explanation's primary purpose *to convince* the explainee to accept the computer's outcomes (perhaps by presenting a simple, plausible, but an unlikely explanation) or it aims to achieve the explainee's literacy about the soundness of the technological solution?

«6» The impact of XAI-related issues (and their collateral effects) are not only computational and commercial in nature (Weld & Bansal, 2019). Systems usability (namely the democratization of their use) on the one hand, and the legal, ethical and social consequences (on the rise in public opinion) on the other hand, play a relevant role in XAI. The unstoppable advance of AI is causing a social and cultural crisis regarding the safety of the outcomes of the systems. Above all, in those touchy realms for society, with relevant media impact (and which are

responsible for a significant part of XAI visibility). For instance, it seems socially unacceptable that an autonomous car does not reach a very high success rate (to safety), close to 100%. Nevertheless, they are far from approaching levels of this magnitude (Biewald, 2016). Not only it is necessary the disengagement decisions -when the human had to take control of a vehicle- are explainable by the technical agents. It also needs to be understandable by stakeholders. Also, it demands that these events do not cause serious accidents. This way, current systems cannot completely ensure that; they are still far from human performance. This barrier suggests the existence of a glass ceiling for autonomous car performance. It is not the only one; it could add another relevant social fact. The best known explanations for public opinion are those related to serious accidents of these autonomous systems. They usually come from diagnostic processes that, due to their importance, seek confidence in the explanation in which all the agents are involved. Thus, it is composed of technical reports, tracing data and normative documentation. The explaining demands are thus actual challenges for the socio-technical system in which autonomous cars are embodied, that require a system-level analysis (that comprises agents such as manufacturers, vendors, institutions) (Borenstein et al., 2019). There exist even more sensitive instances, as the development of lethal autonomous weapons systems. Whether they can determine if a target has military significance (military necessity)? Whether the target is a combatant (distinction)? Whether the military action is overkill to accomplish the task (proportionality)? To date, the U.S. military is worried about whether or not this determination can be carried out by autonomous systems without a human in the "loop" making this kind of decision (Price et al., 2018).

«7» Scenarios of this kind do and will persistently occur. They share an essential feature: the need to tailor the explanation (of CAIS behavior or outcomes) to be acceptable by a layman. Software/hardware is accompanied by human behavior and social institutions in what represents a socio-technical system (Kroes & Verbeek, 2014). This kind of scenario in mind, where *convincing* the explainee (encompassing its psychological dimension) may be more relevant than providing a *correct* explanation. Throughout the paper, it is called an *Explanation-to-Layman-Explainee Scenario* (ELES). ELES can be considered as a complex socio-technical system, compounded by increasing abstraction layers (Fig. 1), some of them hidden by the explainee, although play a relevant role when it comes to the explainer getting the explanation acceptable to the stakeholder. In such circumstances, likely, most of the understanding (e.g. of the software libraries) and the knowledge (of programming and parameter tuning practices) that the explainer deploys to find the solution is not finally represented in the explanation. This is where XAI urgently needs to broach the philosophical, social, and psychological dimensions of the challenge, beyond the pure human-computer interaction, usability issues, and user experience [being aware of the difficulties that exist in trying to reconcile the two fields (Páez, 2009)].

In order to tackle the above challenge, we claim that two ingredients have to play a more relevant role in XAI, which are briefly outlined (and contextualized) below: The Knowledge Level envisioning of AI and the Bounded Rationality paradigm. The following is a brief description of both, but we anticipate the reader that the

Table 1 The roles of knowledge representation Davis et al. (1993)

Role	Description
Surrogate	A substitute for the thing itself. For reasoning about the world rather than taking action in it
Set of ontological commitments	In which terms should the world be thought about?
A fragmentary theory of intelligent reasoning	Expressed in terms of: 1. The conception of reasoning 2. The set of inferences sanctioned 3. The set of inferences recommended
A medium for effective computing	1. An environment in which thinking is accomplished 2. Guidance for organising information
A medium for human expression	A language which says things about the world

paper is not committed to their direct application. Instead, the aim is to take advantage of the ideas and practices of the two fields to outline and assist the XAI process.

1.1 The Knowledge Level

«8» One of the ingredients of the authors's proposal is a proper reading of the Knowledge Level (KL) in Explaining, presented as a way to tackle the problem of building explanations in the three elements: explanans, explanandum and inference link.

«9» Newell's KL begins with the premise that representation and reasoning are intrinsically separated, in the way that inference (e.g. automated reasoning) works with symbolic expressions without intended interpretation (Newell, 1982). The design of the reasoning module concerns only the formal soundness and validity of the reasoning as a mathematical apparatus. Thus one only needs to specify what the agent knows and what its goals are, a logical abstraction separate from implementation details. The behavior comes from the execution of the reasoning module on the representation of the agent's knowledge. From the point of view of this paper, we consider KL-models as surrogate models for rational agents or CAIS (for example, rule-based systems). They appeal in the last instance to the (computational logic) soundness of the modeling. Symbolic models are surrogate insofar as, once checked its soundness, the model satisfies the general requirements of the surrogate ones. Many Data Science solutions starting from mathematical mechanisms (which are assumed to be well implemented in the software libraries) provide support for an inference. From this point of view one could say that even the Decision Layer of Fig. 1 would be susceptible to be modeled for explaining it under KL-inspired principles.

1.1.1 Unlimited Versus Bounded Reasoning in AI

«10» Davis et al. (1993) point out that one of the roles of Knowledge Representation and Reasoning (KRR) is that of being surrogate, substituting the original to reason

Table 2 Rational choice versus Bounded Rationality (extracted from) (Hernandez & Ortega, 2019)

Bounded rationality	Rational choice
Necessity of assistance of the bounded mental capacity of the subject that decides	Unbounded cognitive ability of the subject who decides
Knowledge of an acceptable set of actions	Knowledge of all available actions
Approximate and heterogeneous knowledge of the consequences	Numerical knowledge of all the consequences of actions
Evolutionary and unsettled preferences	Stable and ordered preferences
Temporary and cost limitation that affects the quality of the decision	Unbounded or non-influential resources in the decision-making process
Search for a satisfactory result	Search for the best possible result
Help the one who decides to understand what will happen if he does something	Inform the one who decides about what to do

about the world and infer the decision to be made (Fig. 1). It is not its only role; it is also useful to represent ontological commitments (assimilated to laws of nature) as well as theories for reasoning. KRR represents an environment where both it can organize the information, and the system can *think*. Other approaches as Addis' (2014) (pp. 46) further break down these roles.

«11» Roughly, Rationality comprises five activities: Recognizing and defining a problem or opportunity, search for alternatives to follow, collection and analysis of data on each of the alternatives, evaluation of them in the light of the analysis, and finally, as the result of the latter, the selection and application of the preferred one.

«12» The starting point in KRR would be a scientific background devoted to the study of rational agents. It began with the search for computational models for Unlimited Reasoning (UR), where it was urgent to concentrate on *epistemological adequacy* rather than *heuristic adequacy*. The design of systems under (calculative) rationality arose. Possibly, this leads to inefficient systems, albeit with the hope to get as close as possible to UR using better programming and various approximation and acceleration techniques.

«13» Ideally, the adoption of an UR paradigm -ultimately based on logic inference- would reduce the problem to the choice of sound logic according to the Knowledge Level approach (see below). Heuristic adaptation would still be outside the direct scope of application. We only need insight to choose formalism, expecting that the increasing computational capacity alleviates the problems of the efficiency or the real computability[1]. How to adapt it to a framework with limited resources? There are fundamental differences between pure rational choice and bounded rational choice that should be accounted for (Table 2 summarizes the main differences between both forms of reasoning). These dissimilitudes play a role in scenarios where other rational attitudes come to play as in ELES.

[1] See Dick's paper Dick (2015) for more information on the history and discussions about the origins and difficulties of implementing BR for the first rational agents, and the consideration of heuristics).

1.2 Bounded Rationality as a model for human reasoning

«14» The considerations of human factors as relevant in XAI, beside the need for specification, lead to explore the suitability of approaches based on Bounded Rationality (BR). Simon (1957a) [(see also Simon (1957b)] proposed BR as an alternative basis for the mathematical modelling of decision-making as used in Economics, Political Science and related disciplines. As such, it has to be studied as an indispensable discipline in both, the Social Sciences and AI (and their confluence), a status that it preserves even more importantly today (Gigerenzer & Selten, 2002; Moreira 2019). The basis of Simon's BR lies in the fact that Simon (1957b):

> the capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world-or even for a reasonable approximation to such objective rationality.

Some of the actions we perform which are not the result of a purely rational process are common, due to our intrinsic limitations in the formulation, the processing of information (reception, storage, retrieval, transmission) as well as in the synthesis of solutions itself.

«15» Our ability to work under limitations allow us to face and live within CS that we simply endure or *solve* using incomplete, not purely logical, knowledge (beliefs) reasoning, and yet we are effective at solving or coping with such problems (Duris, 2018). In BR, the decision process is seen -even for relatively simple problems- as a process that does not necessarily choose the optimal action (Hernandez & Ortega, 2019). People's behavior is influenced by both available opportunities and desires, influenced in turn by other factors as their own beliefs. The use of beliefs (which are intentional in nature, and not necessarily true) means that they cannot even distinguish whether some options are viable or not, or whether they are favorable to their interests. That is, the choice could be not necessarily optimal, nor even heuristics-driven (in any case, not necessarily formalized or conscious).

1.3 From Unlimited Rationality to Bounded Resource Reasoning

«16» Let us focus for a moment on resource constraints, one of the pillars of BR, and how this is addressed in KL inspired paradigms such as Agent Theory (AT). Since engineers aim to build feasible engines, even starting from UR in AT, it is necessary to account for some limitations. Among these needs, the cost of processing should always be considered. The (modern textbook) concept of *rational agent* in Russell and Norvig (2003) takes into account the following idea: rational agents try to maximize the utility function according to the resources they have. However, such a definition does not limit the way of obtaining such maximization by logical deliberation.

«17» There exist some approaches, from the classical agent-based AI, to deal with the problem of resource constraints. According to Russell and Subramanian (1995), a *bounded optimal* agent is running the program (from its possible program

space) with limited rational analysis. The authors state that it is necessary to distinguish between two types of costs: the cost of finding the optimal behavior, and that of running the associated program. It is also necessary to point out the distinction between *optimizing the program* and *optimizing the solution that is being obtained*. Russell and Subramanian further clarify the principles that should guide the design of agents under constraints. Namely, bounded optimality is desirable in reality, and it is possible to construct provably bounded optimal programs. Lastly, AI can be usefully characterized as the study of bounded optimality, particularly in the context of complex task environments and reasonably powerful computing devices. Thinking in XAI, the explainer (human or artificial) agents could be driven by these principles. Within the design of systems for XAI, according to the principles of KL, the particular data (perceptions from an event) to be explained should not play any role in the explanation producer (the explaining system). This general principle is observed by the state-of-the-art explaining systems LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017).

1.4 Aims and Structure of the Paper

«18» The paper mainly concerns foundational issues. The aim is to discuss aspects related to the explanation versus the information available (or selected) versus general principles inspired in KRR. We investigate the fundamental issues of the role of KRR in explaining CS behavior. Likewise, the issue of applying BR solutions to achieve acceptable explanations for stakeholders, will be addressed; particularly in the case of ELES within Data Science (e.g. ML technologies and tools for inherently complex problems). Also, throughout the paper, the authors aim to convince the reader that XAI in Data Science owns particular features which should be investigated.

«19» First, we account for the psychological (Sect. 2) and sociological (Sect. 3) dimensions that frame XAI. We point out some considerations on the nature of the explanation as a product from agent interaction, a social construct. Some insights on its impact on the explanation building are discussed.

«20» Second, the question will be framed within another aim that the AI community should accept as indispensable for the promising development of XAI. Namely, the need to incorporate into XAI part of the body of work on Explaining (from Philosophy of Science) and particularly the use of the notion of *mechanism* (Sect. 4). Currently, most of engineering XAI approaches neglect many of these resources that can be useful. A paradoxical oversight, considering that the challenge links to a solid scientific and philosophical tradition -as it will be show in the paper. We do not intend, of course, to make a global review of the extensive literature on the topic, but only to point out some general considerations on the elements of explanation that should be taken advantage of, always within our vision as AI researchers.

«21» The paper is focused on whether one can consider BR for XAI, specifically, within the general question of XAI versus BR (versus logics) The aim is to present it as potential machinery to tackle the *argumentative dialog* driven to achieve the explanation. Starting from (Computational) Logic ideas (Sect. 5), the

Table 3 Features of the system to explain and the model for explaining

On the system to explain	Paragraphs	On the model for explaining	Paragraphs
High Technology	3	Interactive/contrasting	64
CAIS	7	Bounded rationality	Sect. 7
Systems for CS	51	Deductive-nomological	45, 46
		Format	48
		Surrogate	52, 63
		KL	55
		Veracity	Sect. 6.6.2
		Logic limitation rule	99
		BR models	100
		Perspectivism	7.3, 61
		Surrogate KL	53, 51

role of KL-based surrogate models is explored (Sect. 6). The option to formalize the different elements in Explaining will be further explored and detailed in the case of Data Science practices (Sect. 7)

«22» As an ultimate goal, the incorporation of BR in the argument modeling for XAI (Sect. 8), mainly for *local XAI* [(argument modeling is a known resource in XAI (Rago et al., 2021)]. We claim it could be influential for XAI in socio-technical systems for massive data processing. ELES is particularly interesting because circumscribes XAI into expert versus non-expert (e.g., stakeholder) scenarios presenting a considerable knowledge gap. A last section (Sect. 9) is devoted to point out some conclusions as well as future work.

1.4.1 Topics

«23» Throughout the paper, the relationship between some topics and XAI is mentioned. It could be classified according three dimensions of the problem: model/system used, the format of the explanation, and the context in which the work is carried out. The first concerns the system to be explained and the model on which the explanation would be based. (Table 3). The second one is about the three-element sequence ⟨explanans, inference-link, and explanandun⟩ as a format for the explanation (Table 4). And lastly, the use, context, and impact as factors of the socio-technical systems involved in Explaining (Table 5).

Throughout the paper, it will analyze some elements from KL to be applied in XAI (besides BR or isolated in specific sections). Table 6 describes the general approach to XAI from KL, according to the three standard classifications for XAI solutions. In addition, Table 7 lists the commonly used abbreviations used in the paper.

Table 4 Features for elements in Explanations

Explanans	Paragraphs
Mechanisms	38
Boundary Elements	42, 43
Incomplete in CS	95
BR-activity	Sect. 7.2
Variety in BR	Sect. 7.2.1
Inference link	**Paragraphs**
Narrative/story telling	31
Causal vs non-causal	Sect. 4
Causal closure principle	44
non-causal	42
Non-observability	39
BR	Sect. 7
BR logics	7
Limitations of BR	15, 17
Rational	16
Causal BR	113, 36, 114
Explanandun	
Event	
Behavior	
Decisions	

2 On the Role of Psychological Features in Explaining

«24» The relevance of the explainee's literacy is highlighted by considering factors beyond the technological dimension and close to the psychological. Users are more willing to accept automated decisions if the explanation is tailored to their level. Such acceptability refers both to the main features of the domain of discourse where is circumscribed the event/decision, and the use of technological tools, notably to its trust in them (Miller, 2019; Araujo et al., ming). The latter is related in turn to other factors; mainly, to the user recognition or internalization of particular beliefs on the capabilities of such instruments. The internalization could make the explainee accepting the explanation in the same terms as the explainer. This would be the secondary product of the interactive interplay between the two agents that aims to reach a consensus on an explanation. The benefits of internalization are several. When people generate explanations or imagine hypotheses about an event/system's behavior (and thus internalize these), they increase their confidence in those possibilities they have synthesized. Three phenomena would support this (Koehler, 1991): (1) When we use a hypothesis, this benefits from an increase in confidence concerning the rest of the available options for reasoning or argument; (2) When a

Table 5 Context, use and impact of explaining

Context	Paragraphs
Expert vs non-experts	4
Trust	27
Data Science	Sect. 6
Big Data	Sect. 6.4, Sect. 6.7
Absence of Models (BD)	Sect. 6.6
Use	**Paragraphs**
Tailoring	24
Predicting	Sect. 6.5
Taming	62, 95, 108
Curation	59, 60, 112
Semantic Technologies	68,69
Impact	**Paragraphs**
General	6
Acceptability	26, 28
Usefulness vs soundness	Sect. 3.1

Table 6 KL reading of different XAI approaches

Explanation type	
ante-hoc	*post-hoc*
KL-model (e.g. Expert system)	argument-based
Scope on a model	
Global	*Local*
KL based surrogate model	argument-based
General scope	
Model specific	Model agnostic
KL based surrogate model	KL model for explaining

Table 7 Commonly used abbreviations

AI	Artificial Intelligence	BD	Big Data
BR	Bounded rationality	CAIS	Complex Artificial Intelligent Systems
CS	Complex systems	DN	Deductive-Nomological Hempel's explanation
DS	Data science	ELES	*Explanation-to-Layman-Explainee* Scenario
ER	Ecological rationality	KL	Knowledge level
KRR	Knowledge representation and reasoning	ML	Machine learning
SAT	Propositional satisfiability problem	UR	Unlimited reasoning
XAI	Explainable Artificial Intelligence		

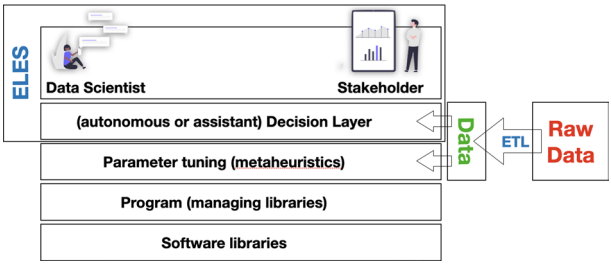

Fig. 1 Stratified architecture on which ELES appears

person is asked to provide an argument (which could be an explanation) to support a certain hypothesis, a collateral effect from that choice (or synthesis process) is that he/she tends to find that argument more plausible than others, and lastly (3) if for some reason, they believe a theory is correct, then they tend to express greater confidence both in its veracity and in the events that will occur from it. There are other factors on the goal itself (explanation acceptance) that could improve the process. Psychologists have determined that some criteria would be a priority to include in an explanation: necessary causes (vs. sufficient), intentional actions (vs. those taken without deliberation), proximal causes (vs. distant), details that allow distinguishing between fact and foil, and abnormal features among others (Weld & Bansal, 2018). The exploitation of such psychological features would facilitate the explainer to convince the explainee in ELES.

«25» Another important issue is the simplicity or minimalism of the explanation, mainly on the representation. According to Lombrozo (2007), humans prefer explanations that are simpler (i.e., contain fewer clauses), more general, and coherent (i.e., consistent with the human's prior beliefs). She also highlights that our desire for simplicity goes so far that we even prefer simple (one clause) explanations to conjunctive explanations—even when the latter is likely to be more accurate than the single clause. This feature supports our idea of working with simple explanation models [(transforming if necessary, relatively more complex logical explanations (Booth et al., 2019)] and try to find its minimalism. One can also take advantage of the study of the so-called *conjunctive explanations*, which are different explanations that, nevertheless, are more explanatory together than separately (Schupbach, 2019).

«26» These disclosures actually link two aims within XAI: *understanding* the event, and *susceptibility to accept* the explanation. According to Dudai and Evers (2014), *understanding* refers to the ability to generate a specific mental model (or a more comprehensive theory) that allows predictions based on the scientific reasoning about the system's behavior. Subrahmanian and Kumar (2017) point out that the term *understanding* is often used in two different ways that do not imply each other. The first refers to the subjective feeling of having a given meaning to something (*we have interpreted it*). The second one refers to having perceived empirical regularities enabling us (subjectively) to predict. In some problems, it is dangerous to confuse them. The former is associated with knowledge whilst the latter could only be the source for solving a ML problem. The second notion could be considered as

the *descriptive understanding*, according to Findl and Suárez (2021). In that paper, the authors study the case of using purely (descriptive) statistical epidemiological models as a tool for decision-making. This kind of model is quite distant from the *explanatory understanding*, the basis for scientific knowledge. It can be argued that the descriptive understanding could not be a solid basis for explaining insofar as they do not offer the necessary epistemological link to the Scientific Theory that supports our knowledge about the phenomena. However, what is undoubted is that this type of model (and similar) which have demonstrated their usefulness and, therefore, it is necessary to study their scope and characteristics (Findl & Suárez, 2021). For instance, investigating whether this kind of model enjoys of what Regt names *Criterion for Understanding Phenomena* and *Criterion for the Intelligibility of Theories* (de Regt, 2017, Chap. 4), and how it can support its potential prediction-generating character.

All this thought is within a framework where notions as *inference* should not be understood as in Computer Science (Computational Logic). Rather, as a human process that does not need to be equivalent to a purely (formal) logical process. A number of non-purely (logical) rational types of information management/processing can be explained by techniques studied in BR, which cover all those reasoning techniques that we use in the face of our processing limitations. Circumscribed to ELES, it would even be necessary to study the effect of the explanation on the explainee's beliefs. It would be necessary to reflect to what extent cognitive biases may affect human understanding of interpretable machine learning models, for example, rule systems. For instance, in Kliegr et al. (2021) the authors summarize them in the particular case of rule-based machine learning models (hence KL-based), pointing out the need for investigating human interpretability from the standpoint of Cognitive Science.

3 On the Sociological Factor

«27» As an inter-agent system, ELES owns a social dimension. One of the factors it has already been pointed out is the explainee's trust in the explanation process itself. Two facets of trust can be studied. One is whether the explainee believes the explanation meets its beliefs about what it is sound explaining. The other one is the inter-agent level: the explainer aims to *convince* the explainee with the explanation. The latter is linked, as it has been argued, to the psychological dimension (and to our processing limitations). According to Cugueró-Escofet and Rosanas-Martí (2019), trust only makes sense in a BR context, where agents are not fully aware of their preferences and values. Trust allows both, the explainer and the explainee, to admit decisions that are not consistent with their beliefs, internal values, or preference systems (something that also could occur due to commercial or political interests, for instance). Trust influences the teleological understanding of Explaining activities, namely the explainee's assumption of the explainer's willingness to act according to the highest values. This is a possible reason to trust that its explanations are of best interest for both agents, even though it may not be the most attractive in terms of the immediate variables of both effort and results (Cugueró-Escofet & Rosanas-Martí,

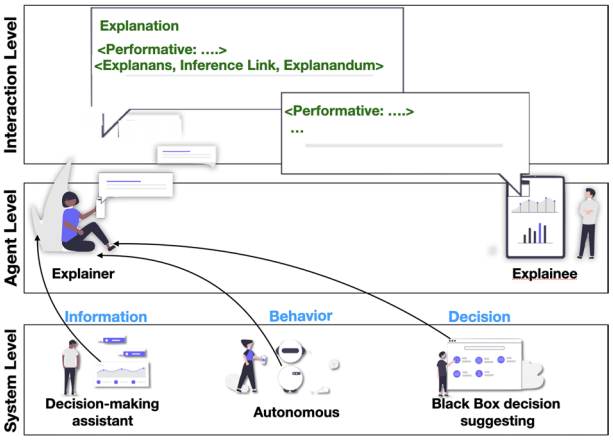

Fig. 2 Socio-technical system where explaining issue is framed. Agents interacts to achieve consensual explanation

2019). At the inter-agent level, the trust induces the explainee's belief that the explainer will make decisions according to his current values (intentions, obligations, objectives, etc.) even when some variables push him in the opposite direction.

«28» Another social factor affecting the success of the explanation in ELES is its dependence on the audience's degree of acceptance of the *narrative* presented. Combined use of several modules in the Data Science Project is time-dependent and susceptible to be explained by describing the trace of the experiments. Thus, there are narrative elements in the part dedicated to the description of the ML model and in the part devoted to justify and argue the decisions made by the CAIS [as for instance within dashboards (Jarke & Macgilchrist, 2021)]. Their existence leads us to explore questions related to the context in which the agents work and the required explanation precision, which is related in turn to the human, social or legal acceptability. The explainer agent should transfer information about model accuracy, using if necessary metaphors to simplify the explanation. For example, employing exemplary models that share the most important variables and values involved in decision making (or the outcome) offered. Also, some strategies based on the adoption of human interpretations allow an excellent balance of performance and intelligibility (Weld & Bansal, 2019) [(see also Janssen et al. (2021)].

«29» Since the paper mainly focuses on *Local Explaining* (i.e., the XAI case in which the explainer agent aims to explain the result offered by the system), the socio-technical system stratifies in tree levels (Fig. 2). The system level concerns the system itself and the outcome to be explained, which depends on the use case: information about the CS, the system's behavior, or the decision suggested by this, among others. The agent level comprises the agents involved in the process, with their associated characteristics. Lastly, the interaction level would be concerned with the documentation of the interaction, its trace, information about the acts of communication, and the final explanation.

3.1 On the balance between usefulness and soundness

«30» The above aspects underpin Lipton's warning in Lipton (2018) on the need to achieve a balance between the effectiveness of models such as Deep Learning and the acceptance by humans of the results they offer. Several factors condition the agreement between agents (the explainee's acceptance of an explanation of the AI-based system behavior) as a social construct. It is supposed that when the explainer interprets the results of the system for the explainee, he/she would be working under the basic principles of interpretive reasoning. The first and foremost is the so-called *Restrictive Principle* (Stern, 2005), that is, only reasonable explainees, who are familiar with the circumstances, understand what is at issue the way the explainer does. The principle shapes the explanation's success but also its usefulness (as reusability).

«31» Also, the restrictive principle seems to claim that the explanation acceptance (and its internalization) also depends on the explainee's ability to develop a (based on belief) mental model of how the tools work and what they actually do, as we have pointed out in Paragraph 24. Hinsen argued that such models are limited by the tool features that we need to know (with the advantages and limitations of that) (Hinsen, 2014). Thus, some of their mechanisms can be hidden or obviated, hence it may exist, to some extent, an undisclosable view of the mechanism/tool (Goebel et al., 2018). An extreme case appears when the system users are unable to make an informed decision between different models to choose the most convenient or efficient program, regardless of which model it implements.

«32» Other extreme scenario occurs when the explainer focuses on the goal of explainee's acceptance, since the system is provably correct but the stakeholders demand an explanation before accepting the decision the system suggests. Among other options, *fictionalization* of the explanation (or the model) can contribute to the success. For example, *Storytelling* strategies that exploit metaphors associating explanations with explanatory traditions from other sciences are shown to be useful. The advantage of this kind of strategies is that they are focused on convincing the explainee, and therefore there is a certain relaxation of the *completeness* of the narrative/explanation. This approach is actually producing a *narrative*, not a truly explanation itself (compelling versus scientific soundness). Thus, the *incompleteness* of properties among *fictional entities* (those used to mount the metaphor) is not a simple anomaly (Margolis, 1983), actually is *part* of the strategy itself. A consequence could be that the explanation becomes doubtful for different explainees. Thus, the reason for the non-preservability of the validity would be its ontological status of the so-called *embedded narrative*, which are mental representations, produced by a history, that are virtual. They are not verified in the factual domain, being thus epistemologically weak insofar because they belong to the mental/subjective realm; they are susceptible to reinterpretation or transformation by another explainee. Explanation malleability represents a new source of risk. Finally, the uncontrolled modification of the explanation (and its practical consequences) across the organization can exacerbate the bad practices that XAI aims to prevent, as it already occurs in the Privacy field in DS.

4 Causal Versus Non-causal Explanations

«33» The complex relationship between logic -entailment- and causality has been largely studied in Philosophy of Science. Therefore, it is not surprising that dilemmas as causal versus non-causal explanations remain in force within XAI, even if it is blurred by the requirements of explainee's acceptability of the explanation.

«34» Hobbes claimed that a phenomenon is explained when one assigns a cause to it. Knowing the causes of an event (in some of the various and disputed meanings of causation) is considered one of the most solid forms of explanation. Sometimes even more important than proper understanding of the event itself. The status of causality in XAI is possibly due, to the preponderant role that Physics has played and plays in Science in general, and therefore in Engineering. So much that mechanisms and causality are very useful resources in the enterprise of explaining and justifying events, theories and results. Also, the interplay between Philosophy and Science matters.

«35» This section does not intend to be a general review of the features and the causal versus non-causal dilemma in Explaining which are inherited in turn from Science [Miller's (2019) contains a general discussion of the topic]. However, some notes on the role of causality in CAIS are necessary to frame up the following sections (several hard open problems in AI are intrinsically related to causality). Mainly, its relationship with the notion of mechanism (from Philosophy of Science and Physics), to the extent that many CAIS can be broken down into what can be considered their basic mechanisms. This idea is to support some arguments developed throughout the paper, such as our claim that *logical mechanism* (not understood here in a narrow sense) has sense in XAI because it shares some of the features with the classical notion of mechanism in Explaining. Although it could be far from a logical-computational paradigm, it is interesting to consider certain logical and mathematical steps as mechanisms, or at least to consider that they admit such a reading. Of course, this consideration is not free of controversy, since not all current trends in Explaining Research admit an identification of explanation with argument, and logic usually produces the latter (Huneman, 2018). Moreover, the mechanistic conception of Explaining usually breaks with the idea of explanation as entailment, which allows us to avoid some classic critics on this idea (Huneman, 2018). Although it is not our aim to rely on the mechanistic view for the discussion of *logical mechanisms*, we do believe it is necessary to devote some space to it. Mainly, due to three reasons that arise when we focus on ELES. The first is that ideas of Mechanicism are implicit in certain practices of ML researchers for explaining. We refer, for example, to those who analyze the numerical interaction of nodes or subnetworks within the whole network as mechanisms that, combined, explain the system. However, such explanation is not useful for the explainee (since it could be uninterpretable in practice), despite it could be a valid one from a mathematical point of view. Secondly, the observation mechanism does not imply achieving the understanding of the explainee. And lastly, these warnings may be valid even for the logical mechanisms, which makes their study necessary.

«36» Thinking about the problem in ELES, one has to recall how humans usually can recognize causation. One form consists of comparing (could be mental) the outcome when an action is taken with the corresponding when the action under study is retained. If the two results differ, we say that the action has a causal or preventive effect on the result. Otherwise, we say that the action does not have a causal effect on the outcome. The idea also fits with Craver's notion of the variable *causally relevant* (Craver, 2007): a variable X is causally relevant to the variable Y in the conditions W if any intervention on X in the conditions W changes the value of Y (or the probability distribution over the possible values of Y) (Craver, 2007; Barberis, 2012). In ELES, the causation problem could exacerbate because the explainee should understand that the difference between the outcomes makes causal the action or element. Also, in its understanding of the role of the features involved. The resource of considering *mechanisms* (actual or as metaphors for systems subprocesses or modules) could alleviate the knowledge gap.

4.1 On the Role of Mechanisms

«37» It is often to consider causality as a *mechanistic tool* for explanation in Science, within the broad consensus in Philosophy of Science about the soundness of *mechanistic conception of the explanation* (with reasonable discrepancies and evident weaknesses). According to such conception, to explain a phenomenon consists of displaying the relevant parts, activities, and organizational features of the mechanisms in which that phenomenon has taken place. Hereby, the searching for an explanation would focus on the searching for *mechanisms* that, combined in a certain way, will produce a final effect of the observed event. Notwithstanding, the notion of the mechanism itself may be subject to discussion. Chiefly, the level description of them can lie anywhere on a continuum from a mechanism sketch to an ideally complete description (Craver, 2006). An important observation to be taken into account is that mechanicism does not focus exclusively on the etiological explanations of the event, but rather on constitutive or component explanations and its representation itself. It would be a purely epistemic approach in that sense, opposed to the *ontic* conception of the explanation as an object independent of its representation (Salmon & Press, 1984).

«38» But what does mechanism means here? Since different proposals exist, instead of embracing a particular definition, it is interesting for the paper to adopt Hedström and Ylikoski's vision (Hedström & Ylikoski, 2010). They argue that a mechanism can be usually identified with the kind of effect or phenomenon it produces; *a mechanism is always a mechanism for something* (Darden, 2006). In the authors' sense, a mechanism is an irreducibly causal notion. It refers to the entities of a causal process that produces the effect of interest. It is also necessary to take into account whether this mechanism is observable by the two agents involved (explainer and explainee) or rather, what degree of disclosure does it show to them- since such a point would affect XAI problem.

4.2 On Mechanism Observability: Undisclosable or Explainable Ingredients

«39» The non-observability of the mechanism by some of the agents can be quite controversial. Options such as emergence-based explanations in Agent-Based Modeling of CS may suffer non-observability to some extent, producing what we could call *epistemic gaps*. We do not only refer here to statistical-computational mechanisms but to certain intrinsic inaccessibility of the mechanism, such as those (inaccessible links) that connect micro and macro levels in emergence phenomena in CS and particularly in complex neural networks.

«40» Observability is linked to another characteristic feature of a mechanism, according to Hedström and Ylikoski: *it has a structure*. It should be possible to disclose it, making visible how the participating entities and their properties, activities, and relations, produce the effect of interest. For example, the focus on some subnetwork of a complex neural network would allow the designer to understand its role/function within the overall system. However, embracing a KRR standpoint, a black box ML system might be undisclosed. Even opening it, its actual logic-mathematical structure difficult to understand by the layman its behavior (according to the Ihde's High Technology notion). It can be concluded that, if it is required that any mechanism involved in the explanation can exhibit its structure, then some complex logic mechanisms (as some modules within state-of-the-art SAT solvers) would not be considered as such (if one adopts the vision of the authors). Finally, another feature is that the mechanism does not have to use only explainable ingredients. In fact, to build an explanation one can use non-explainable ingredients such as fundamental principles, elements that would be *boundaries* or *limit*. They would be considered explainable *per se*, or *nomologic*.

4.3 Boundary Elements Versus Causality

«41» In Scientific Fields such as Physics, there is a growing tendency to propose non-causal models. This type of model poses serious foundational problems, because there would be a certain need to outline sound conditions to decide whether an explanation is acceptable or not. By neglecting causality there is a risk of presenting models that provide *only* non-causal explanations. The risk is present even in state-of-the-art explaining systems as LIME (Ribeiro et al., 2016) or SHAP (Lundberg & Lee, 2017), in which the role of some system features are drawn employing statistical or game-theoretical tools, being causal factors hidden for the user. This type of system helps the engineer to capture the generalizable patterns underlying the outputs of a system. Such patterns allow to make inferences about the (potentially causal) connections between the inputs and outputs of the system. It would be necessary to discuss how to enrich these approaches to provide more convincing information from the explainee's viewpoint (including some representation of causality, if necessary) (Pearl, 2009; van der Waa et al., 2021), or techniques for emergent semantics [(e.g. (Borrego-Díaz & Páez, 2022)].

«42» To analyze the issue in XAI -independently of the explaining model-, it is necessary to return to the Philosophy of Science. King (2020) proposes two conditions upon which to base the analysis of the model for XAI. The first is the Local Counterfactual Condition (LCC):

> An explanatory model M provides counterfactual information that shows how the explanandum E depends on M and initial, boundary, and auxiliary conditions C.

The second King's condition is called the Global Confirmation Addition (GCA):

> An explanatory model M is a part of or may be in accordance with, a highly confirmed scientific theory T.

There exists the risk of considering GCA as inaccurate, although GCA could actually point to the nomologic part of the explanation. It needs an accurate analysis of what the three notions in GCA actually mean within the technological realm of XAI: *to be a part of*, *to be in accordance with* and *a highly confirmed theory* (King, 2020). Continuing the parallel with Physics, some new models have such a degree of abstraction and epistemological richness that there may be conflicts in the model description itself. This circumstance casts doubts on their usefulness. Overcoming the obvious distance, some CAIS may provoke similar doubts. CAIS not only are nonsymbolic systems as complex neural networks, other systems as state-of-the-art SAT solvers also fall in this category (Giráldez-Cru & Levy, 2016). Nomologic aspects of explanations in the latter should be logical in nature, but not only that. Others can represent data specifications, well-established algorithm schemes, etc.

«43» The assumption of the existence of limit/boundary ingredients in explanatory mechanisms (or in the explanations in general) could serve to outline a frontier between causality and non-causality (Sullivan, 2019). Reutlinger (2014) defines a non-causal explanation (NC) as one that contains at least one non-causal element *e*, and in addition, *e ensures the success* of the explanation. The notion thus expressed would be problematic in XAI. If *ensures* means that it is a condition for an explanation, then NC is too exclusive, since any explanation that includes at least one non-causal element would be NC. We should accept the existence and use of certain boundary conditions (boundary elements) even if they are not causal in nature. Boundary conditions are necessary to construct the explanation, although no information is available about their causes (in fact, one would admit it is not actually necessary). In any case, by continually reducing ourselves to causes we would come across some *base* or *limit* ingredient that will be inherently non-causal. In CAIS, such elements would be elements of confirmed or well-established AI theories, or simple transformations of inputs (from sensors, for example). The most evident example will be the raw data provided by the perception or I/O modules. Non-causal ingredients elaborated from this kind of data could be a standpoint of the world representation (expressing restrictions on the features on which the explanation is based).

«44» The argument developed so far in the section seems to lead to a *physicalist* point of view of the explanation. For example, a backtracking analysis of

causes leads to consider principles as the so-called *Causal Closure Principle* (CCP): *If a physical effect has a cause, then it has a sufficient physical cause* (Dimitrijević, 2019; Papineau, 2001; Kim, 2005). This may seem an extreme position, but would be necessary in safety critical software, for example. Acceptance of CCP as a requirement would require a rigorous treatment of all the elements involved from a mathematical, logical-computational standpoint (which could be counterproductive to the explainee's acceptance of the explanation). In any case, one would have to ask oneself which are the boundary elements and whether they would represent the *laws of nature* of the specific Computer Science or XAI ground theory, or even consensual beliefs in ELES. Notwithstanding, these kinds of principles can sometimes be hidden by the causal roles. This seems to be one of the more controversial principles on Explaining practices in the *new mechanicism* (Huneman, 2018). If we want to take as much advantage as possible of the analogy between mechanism and logical mechanism, then it is necessary to determine such laws. Additionally, the boundary elements, concerning causality in the system to explain, should be addressed. Both aims could greatly help the design of models for explaining beyond the statistical ones.

4.4 Laws of Nature and the Structure of Logic-Based Explanations

«45» The consideration of the *Laws of Nature* as ground ingredients for explanations is ubiquitous in Philosophy of Science. For instance, examine us the so-called Deductive-Nomological (DN) Hempel's concept explanation (Hempel, 1970). The first condition is that an explanation is an argument or inference equipped with propositions for premises and conclusions and the relationship between both (which is the expectation of obtaining the conclusion based on legal connections). Hempel's second condition states that explanans must contain at least one law of nature (the nomological component) so that the derivation of the explaining would not be valid if it removes the premise. One might wonder how these laws would be when one wishes to transfer the requirements to XAI.

«46» The *KL laws of nature* can emerge if the explainer imposes some kind of minimalism on the elements of the explanation and grounding on the background knowledge, where such laws may already belong to. These laws of the domain where the CAIS applies, would be represented within the background knowledge. This way, Hempel's condition is satisfied.

«47» Even so, following the DN model leads to inherit its associated issues, such as relevance and asymmetry (Woodward, 2019). Other authors also concern with pragmatics, as van Frassen Bas (1980), who defends the pragmatic and intentional view of explanation. Among other proposals, he relies on the purely logical idea of explanation, although this relation is understood here as from question to why-question, and the answer would be an essentially contrastive explanation. There exist also other proposals to solve the difficulties. For example, by extending the explanandum domain specification to represent richer representation frames. [cf. Díez (2014)].

5 Grounding on (Computational) Logic Principles

«48» At a logical level, the explainer can exploit the implication as causality relation (although sometimes actually represents correlation). It is a well known and popular option with a longstanding history in AI (e.g. in Expert Systems). and represents a successful approach to the problem.

«49» The adoption of a paradigm from Computational Logic does not exempt us from encountering difficulties. The first is whether the right approach has been chosen. One of the risks that should be avoided when building explanatory models is the so-called *Heuristic Fallacy (HF)* (Gabbay & Woods, 2003):

> Let H be a body of heuristics with respect to the construction of some theory T. If P is a belief from H which is indispensable to the construction of T, then the unqualified inference that T is incomplete, unless it sanctions the derivation of P, is a fallacy.

A Computer Science vision would be to claim that HF deals -in XAI- with the problematic relationship between the approximation to a theory and its applicability. Accepting the belief *P* is indispensable is very useful to ensure the explanation (considered at one time as a specific theory) is acceptable. As Gabbay and Woods aim higher (Gabbay & Woods, 2003); they argue that if the theorist avoids the fallacy, then it is likely that the procedures derived from the designed theory could be inapplicable (for example, due to computational complexity). Therefore, one could claim that theorists must avoid HF but, at the same time, adjust the theoretical models in a rough way that they have realistic executions.

In ELES, another complexity ingredient comes from the difficulty of translating some logical features of the explanations into a language acceptable and intelligible by a non-expert. Two elements need to be translated frequently. The first, of course, the explanandum but also the second: the part of the Knowledge Base (KB) used to entail it (boundary elements, nomological components or laws of nature). That is, the initial hypothesis besides the inference links. Regarding the explanation link, it should also be translated or adapted when it is not comprehensible to the explainee (Booth et al., 2019).

5.1 Explaining by Using Knowledge Level-Based Surrogate Models

«50» The undertaking of building models to support explanation -especially for intelligent systems based on ML- covers various techniques, ranging from those specialized in Deep Learning (cf. Townsend et al., 2019) to logical causal models (in the tradition of classical Expert Systems). To approach the issue from KL, the need to reconcile two levels of reasoning (the explainer's and explainee's) through some accepted (consensual) model, becomes more pressing. What Newell's Knowledge Level (KL) (Newell, 1982) paradigm can offer to meet the Explaining challenge is mainly explanation, interpretation, and justification. These are research practices deeply rooted in AI, as they provide reliability in autonomous systems for the

decision-making process. However, what would be the status of a KL-based model within XAI? Can one consider this as a *surrogate*?

«51» The standpoint is the simplest and currently most common use of surrogate models, namely, to obtain results at less (computational) cost than those obtained through expensive (actual) experiments. The basic idea is that the surrogate model acts as a *curve fit* system to the available data so that it can predict results without resorting to the original. At a higher level of abstraction -and thinking on CS- the idea of building a surrogate model is very appealing.

«52» Surrogate models (cf. Forrester et al., 2008) represent a common approach in Engineering. It is a natural approach when attempting to understand the system behavior or to explain a complex event. For instance, the model can approximate the system/event behavior under certain guidelines of rationality, allowing the engineer to work with the model to achieve plausible explanations of the original. An example of a surrogate model to the explanation itself is LIME (Ribeiro et al., 2016). Focusing on general XAI, one should consider some nuances since the problem also aims to preserve correctness. Explaining a CAIS -its outcome or behavior- by working with approximate models blurs the border between the errors present in the original system and those produced by the approximative nature of the surrogate model itself. In other words, the source of explanations can be at the same time the source of new errors or misunderstandings due to the *granular view* of the system's behavior which the surrogate model represents. Of course, that issue is not specific to XAI (it also occurs in Agent-Based Modeling of CS), although it is true that representational fidelity is decisive here.

«53» In contrast to the usual surrogate models in Engineering, within the KL the agents or systems mainly work with variants of symbolic (logic) reasoning. They seek representing information from the environment to obtain conclusions by means of mechanized symbolic manipulations, without any intended meaning. In this way, it is only necessary to specify agent knowledge, beliefs, and goals. When considering KL-based surrogate models, it is necessary to assume the separation between the logical abstraction and the algorithms and implementation details of the inference/decision process itself. The separation of representation and reasoning modules aims to study without ambiguity the KRR's own problems in a separated way: representation and reasoning.

«54» Although KL-based models, as the rule-based ones, represent a sound solution for XAI, in general, providing a KL model for explaining does not necessarily solve the problem itself. This could present a complex behavior that is hard to both specify and prove its correctness. The logic exploited in the model is not necessarily helpful to explain the output. KRR technologies, those used in the internal level (on data) and the external one (on the system), can be *ontologically separated*, that is they might be based on principles that are distinct or uninterpretable between them (e.g. Evans et al. (2021)). This issue is frequent while working with the already mentioned emergence-based explanations, where the inference link can end up hidden within the ontological gap between description levels.

«55» Within the KL perspective, the models providing outcomes closer to the explainees' reasoning practices would be more likely to be accepted. Thus, the explainer will aim to synthesize an explanation similar to that a human would

provide. If one wants to build KL models for surrogating, it will be necessary a proper reading of the factors that make surrogate models in Engineering (quantitative in nature) useful. Roughly, these come from explainable model training on the original inputs and predictions of the complex model. Here we refer to explainable models for stakeholder, such as linear regression or decision trees, which are accepted as explainable in the Engineering Community. As the classic surrogate models are useful for explaining non-linear, non-monotonous models, the ideal KL-based surrogate model would rely on basic inference mechanisms to explain the event, the explanandum. It is in this aspect where authors think that BR solutions could aid building KL models. Another aspect that reinforces this position is that the KL-based models can provide natural solutions to the interactive phase preceding the explanation acceptation, by providing arguments from system behavior, as for example, what DARPA proposes (DARPA, 2016).

6 Explaining in Data Science. Curation and Perspectivism

«56» The assistance of data scientists to stakeholders studying/taming the behavior of a CS (for example, explaining or assisting in particular decision-making on a social network) raises the question about what would be a plausible explanation acceptable to the latter. For example, General CS -such as Urban Mobility Systems or Smartgrids- have multiple levels, which are *open*, that is, they are influenced by the *outside* and interact with it. Thus, an explanation focusing on a restrictive view does not necessarily provide the *best answer* (or even a correct answer). Therefore, it is reasonable to expect that this would not offer general solutions to the explaining problem. The alternative path to mitigate the deficiency would be to address its adequacy from the *perspective* data scientists are led to by the selective access to massive data, as well as regarding the inevitable biased selection of dimensions, features, and the datasets themselves. That is the perspective that emerges from the data, from its curation and exploration, compromising the desired independence of the observer's view. Such a standpoint -underpinning the human factor that configures the perspective- is a particular instance of more general concern. In research areas such as human cognition and quantum physics, traditional science is being questioned as an independent observer approach, inspiring a *Second-Order Science* (SOS) (Müller & Riegler, 2014) that analyzes the challenge from a meta-science level.

«57» Perspectivism emerges from the premise that all perception and ideation take place from a particular perspective (i.e., from a particular cognitive point of view). It is assumed the existence of different conceptual schemes -from perspectives- which influence how the phenomenon will be understood, as well as the judgment of its veracity. Although it is assumed that there is no single *true* perspective to explain the world, it is not supposed that all perspectives are equally valid. In a perspectivist view, Science would be primarily observer-dependent. We are witnessing a growing recognition in scientific studies that most of scientific knowledge is perspectival (Alrøe & Noe, 2014). The context established from a scientific discipline is decisive for the kind of observations, and hence the results. If it transfers the idea to

a more modest environment, the same ground phenomenon occurs *intra* theory, that is, different contextual observations that ground on the same theory.

«58» What is claimed here is that the explanations are inherently perspectivist artifacts when working with CS in ELES. From an *inter*-theory standpoint, Perspectivism claims the existence of different scientific perspectives to analyze a complex problem, all of which can bring value to the study, knowing that in the case of general CS, a single scientific discipline can not provide sound solutions to their complex behavior. This limitation reinforces the role of scientific models in the *absence of models* controversy in Data Science (Sect. 6.6 below).

6.1 Towards Perspectivism in Data Science. Curation

«59» The challenge of the fast development of our digital ecosystem confronts us with data and information on extremely complex problems. Internet provides an astonishing amount of information about any kind of phenomena and events (social, humanitarian, economic crises, etc.). Within the DS universe, an interesting and valuable kind of information comes from the effects themselves of important socio-technological problems. Paradoxically, although this could help to understand and manage them, the wealth of data on the event could prevent its use. It is unavoidable to select, *curate*, data. Consequently, the solution proposed for the problem will depend on how data scientist has approached it (profiling the starting conditions and the concept of solution) and vice versa. The *ad hoc* parameters -with which will evaluate the soundness of the solution- are also specified. Once the aim has been profiled, data scientists resume the data curation practices [This stage would represent the *foraging loop* in the sensemaking process (Pirolli & Card, 2005)]. Therefore, the definition of the problem ultimately will depend on the sketch of the solution (and thus on the explanation as well). This loop could make challenging obtaining a proper formal specification of the problem in BD , which leads to opt for descriptions of some similar *well-behaved* problem that scientists could solve, and to claim that this is the problem to solve (Rittel & Webber, 1973). The problem has been extensively studied and contextualized in other domain fields [e.g. Leonelli (2016)].

«60» One could reasonably conclude that this kind of practice conforms to an actual perspectivist approach to data processing in BD. That is, a *perspectival Data Science* that can be interpreted as the translation to DS of the sensemaking in Intelligent Systems (Klein et al., 2006). In DARPA (2016), the DARPA agency motivates the Explainability Challenge approach in Data Science with -among other arguments- that decisions assisted by DB analytics need a selection of which resources will be the target of study to support evidence in the analysis. Curation itself may induce failures or errors that need to be analyzed, to refine both the procedure and the content curation. The provision of effective explanations obtained from robustly curated data would greatly aid XAI DARPA (2016). From a Philosophical standpoint, one can dare to suggest that what is proposed by DARPA goes beyond solving the challenges of data curation. The agency seem to point out the need for the design

and practice of *data hermeneutics* (Romele et al., 2020; Gerbaudo, 2020), which would cover the overall process, the result and the extraction and curation policies.

«61» In spite of the problems, data scientists perform data curation even for finding explanations. Moreover, they actually curate based on BR practices. While exploring data, the explainer uses intuition or heuristics, like in other processes of Knowledge Management [for which several features have been identified (Hvorecký et al., 2013)]. Since several of such skills affect the results, the overall outcomes should be considered as the product of a certain perspective, built to explain the system's behavior in the philosophical sense (see Sect. 7.3 below). The existence of several explanations (supported by different theories, experiments, tools, concepts, or categories) supporting a CAIS-based decision leads us to move the question of what would be a *correct explanation*. The question reminds us of the persistent challenge on how to work with a *non-unifiable plurality* of partial knowledge (Longino, 2006) [see also (Alrøe & Noe, 2014)].

«62» Turning back to ELES, another circumstance that strengthens the idea of explaining as an outcome of perspectival DS practices is that many BD problems come from well-known classical social or technological phenomena. Data are useful, for example, when scientists aim to engineer a socio-technical system producing similar data from its observed performance (Jones et al., 2013). But this information also becomes a new ingredient to be used while tackling a kind of old and known challenges [e.g. the *wicked problems* (Rittel & Webber, 1973)]. The nature of this kind of information suggests some interesting questions beyond XAI: Can AI aid experts to *tame* these problems on which they have a large amount of data? Is it possible to address the problem by reasoning with knowledge extracted from that source? They are two relevant questions because technological solutions supporting affirmative answers have to be explained. For ELES, the AI tools can provide (statistical, logical, or other) support to decisions or information received, although with particular nuances. Here the explanation of the decision does not aim to convince of its correctness, only of its satisfiability.

6.2 What Does the Explainee's Understanding Role Play in Data Science?

«63» It has been argued that cognitive factors play a relevant role in the explainee's acceptability of the model. In ELES, it is necessary to keep in mind that a relaxation of the -logical, mathematical, or statistical- requirements should not lead to the emergence of misconceptions, as statistical fallacies concealed within High Technology. It is a risk the case of considering as significant a particular phenomenon when a large amount of data is available. For example, *Gambler's fallacy*: if something has happened more frequently than usual, then it is now less likely to happen in the future. Techniques based on Bonferroni Principle can help to identify such (random) occurrences and avoid treating them as an actual phenomena. In addition, during the data exploration phase, the data scientist can decide to focus on particular feature sets and study the relationships among them. There are many more risks, different *illusions of validity* associated with the processing and exploration of data (cf. Aronson, 2011), chapter 2). This leads to the necessity of reconsidering both the

features and data dependencies (Gajdoš & Snášel, 2014). These risks are shared by the explainer and the explainee agents.

«64» Even using intelligible models for the stakeholder, it is likely that the process of explaining remains being an interactive and contrasting task, something deeply analyzed in XAI (Stepin et al., 2021). It involves questions such as *What happens if a condition is altered or eliminated?* or *What happens if the condition used in the explanation does not occur?*. The explainee could also ask for different model instances (even different models). Lieto, Lebiere and Oltramari (Lieto et al., 2018) refer to another two problems, common to most cognitive architectures (CA), affecting their representation level: the limited size and homogeneous typology of coded and processed knowledge. Such issues would be inherited in perspectival DS. It is worth noting that they are not purely technological problems, but also epistemological in nature. For example, they could limit the plausibility of comparing mechanisms of representation and processing of knowledge with those executed by humans in their daily activities.

6.3 The Data Scientist Within the Loop

«65» An extreme case of perspectival DS is that of the promising *citizen data science* projects, in which citizens will handle auto-ML systems (e.g. implemented in their smartphones). The user/observer would be embedded in the context itself, hence the observation and data sources would be curated by she/him. Such circularity should be taken into consideration since it can even affect the desired causality (Füllsack, 2014). Something similar occurs in XAI for Neuroscience, in particular, when assisting in the closed-loop approach to treatments Fellous et al. (2019) or in neurostimulation, (Fellous et al., 2019). Another similar case of an *embodied explainer* would be the challenge of self-explanatory machines, for instance, those with self-diagnosys capabilities. In the event of an incident, an autonomous car with diagnostic capacity would check whether it is your responsibility (leaving aside for the moment the supposition that the idea of a full autonomy obviating the need for human-machine collaboration is very arguable (Bradshaw et al., 2013)). For carrying out the task, the agent must work within a very complex system of responsibilities relationships, and role-taking modules (Kridalukmana et al., 2020). Self-diagnosis becomes a true challenge since the agent is located in the environment where data are recollected, hence it can influence this. Recently, the National Transportation Safety Board Office of Public Affairs (NTSB) provided factual information via a public docket for two Tesla accident investigations.[2] Part of the information is retrieved from the vehicle, which represents the ground documentation to synthesize the explanation (the diagnosis of the accident). In this case, there exists the requirement that the diagnosis must be not only rigorous; it must also be intelligible to political and business leaders (as stakeholders they are in an ELES), beyond the

[2] https://www.ntsb.gov/news/press-releases/Pages/NR20200211.aspx.

simplified explanation to be provided to public media.[3] Another factor that is not discussed in the paper -and that plays its role- in the case of unmanned vehicles, is the compulsory requirement to offer morally acceptable explanations from machines (with morality learned from ourselves (Awad et al., 2020)). Such a requirement is only a particular case of the general challenge of deciding whether the system *has embedded* some values, such as fairness, transparency, explainability, and accountability (van de Poel, 2020). Their compliance can, in turn, be a source of the new explanation needs.

«66» Some sort of similar circularity occurs in monitoring, personalization, or recommender systems. Their daily use involves explainees (clients or users) who observe and confront his observations with the explanation offered by the explainer. The observations -and their evaluation in the light of the explanation- may be affected by biases (in fact, the manipulation of the model leads to accepting biased models, one of the vulnerabilities of XAI solutions (Slack et al., 2021)). Also, other limitations such as slowness, imprecision, subjectivity, and the need for granularity -which are typical of human perception and cognition (Anderson & Perona, 2014)- arise.

«67» Working within a massive data framework exacerbates some of the above issues. In BD, data scientist are dealing with software that needs to manage thousands of features [among other issues (Li & Liu, 2017)]. Moreover, performance requirements are likely to force the adoption of methods that are notoriously difficult (or impossible) to explain; unconscious human skills play a relevant role. This is the case with complex deep neural networks or enhanced decision forests (Weld & Bansal, 2019). It is often the case of *post-hoc* explanations may be the only way to facilitate human understanding.

6.4 On the Role of Semantics

«68» Focusing on the potential use of the KL ideas in DS (which could be considered a limited case of technical explainability), problems of similar magnitude arise. Even focusing on the Semantic Web envision or the application of semantic technologies, one faces the classic KL challenges but with a greater dimension. The treatment of inconsistent/incompatible features is only one example among many difficulties (Alonso-Jiménez et al., 2006). There are other similarly complex problems, for example, issues related to incompleteness, or those associated with the complexity of the involved ontologies that became in actual standards as Genetic Ontology GO in Biotechnology. Another issue would be the lack of relevant metadata, which is unknown and unknowable due to the impossibility of inferencing them employing some intelligent method (a typical case when working with knowledge graphs).

«69» The incorporation of Semantic Technologies into massive data analysis -such as those applied to knowledge graphs (Nickel et al., 2016) is promoting AI systems that deal with elements closer to the user's mental models than the purely

[3] See the link: Data shows Tesla owner experienced repeated glitch days before deadly 2018 crash

numerical ones. Even it could offer complementary information about the result that could serve as an approach to explaining. In this case, powerful tools to represent the *reasoning followed by the algorithm* -closer to the explainee literacy- would be available (Wang et al., 2019) (Borrego-Díaz & Chávez-González, 2006; Aranda-Corral & Borrego-Díaz, 2010). Furthermore, rather than attempting to confirm the explanation through purely deductive approaches, semantic resources such as *Linked Data* can facilitate the search and analysis of counterfactuals (Janssen & Kuk, 2016), rather than simply collecting a representative sample of data to confirm our theory.

6.5 Explaining and Predicting

«70» So far, the section has mainly focused on factors concerning the quality of the explainer's argument (i.e., the explanation). We have claimed that these can be variable, context-dependent, and driven by ultimate aims, but their reusability has not been mentioned yet. Often the aim is not only to explain other events but also to provide an aid to predicting. A *post-hoc* explanation if this does not enjoy some *predictive usefulness*, could not be enough when facing similar events. The scale of the problem is once again a determining factor that conditions such an expectation.

«71» The emphasis on *prediction from learning* is what actually endows meaning and utility to several AI-based solutions. It is assumed that the knowledge (belief) recovered must allow making meaningful predictions; it is not enough to explain what happened. The requirement should be similar to that which Karl Popper proposed for scientific theories. By demanding *meaningful* predictions, we are implicitly admitting that experiments or scenarios can be put forward that call into question the causes and explanations. In this way, we strengthen our model through the contrastive explanation of the phenomena.

«72» The enormous empirical (and theoretical) uncertainty in massive data processing tends to overwhelming attempts at reliable prediction in many socio-technological realms. Its use to talk about the future is limited by foundational (teleological) issues, and it should be so to adhere to best practices. In these cases, predictive modeling may be more useful as a heuristic tool for generating possible scenarios than as a producer of specific policy advice in ELES. In other fields as Computational Social Science, researchers urge to combine explanation and prediction in order to tackle data challenges (Hofman et al., 2021).

«73» Popper's requirements admit a particular reading in the case of massive data. The vertiginous advance in algorithms and technology in DS opens a significant gap between the safety of Science and experimental results on the one hand, and the use of algorithms (considered as) valid or useful on the other. For instance, there is a new need of analyzing the *sensitivity* of the inference in BD to changes in the initial hypotheses, to understand the degree of *robustness* of the results (either decisions or explanations) concerning certain features. Also, in BD, the problem of *causality from data* worsening due to -among others- its multidisciplinary nature (Wong, 2020).The *contrastive dialogue* among DS and Scientific Theories is not an easy undertaking. One of the reasons is the arguable role that part of the DS community endorses in the domain theory, that is, the scientific counterpart of DS practices.

6.6 The Absence of Models and Purely Empiricism-Based Explanations

«74» Although the requirements for the explanation could be mandatory for our confidence in critical AI-based systems, one should not expect to elucidate, with these artifacts, the confusion of information about an object with knowledge about this. The issue has proven itself to be very dangerous for our society today. The Scientific Community is very attentive and concerned about the challenges inherent to data processing and the impressive deployment of AI techniques. In a BD context, there are additional problems that come from the hype itself. One of them is the overestimation of the capabilities of High Technology that is grounded neither on a universally accepted explanation of the decisions (or the solutions provided by massive data processing systems) nor on the existence of a scientific or social model to support the explanation displayed by the AI engineer. The overestimation is an actual social belief fed by Social Media that distorts a proper confidence level in such systems.

«75» However, these are not the only issues. Sometimes the social belief in the validity and usefulness of the system is wrongly supported by the large volume and heterogeneity of the data it can digest. The *Volume*, as an isolated dimension, does not characterize BD itself. It is necessary that the *scale jump* also affects other dimensions such as *Velocity* and *Variety*, in such a way that a change of paradigm is needed because the classical solutions are not suitable. We are also facing new problems related to *Veracity* issues as the so-called *absence of models*.

«76» In 2008, Chris Anderson, editor of Wired magazine, published an article on the *data tsunami* and Science (Anderson & Perona, 2014), within a special issue on DS in the face of the huge amount of data that was already flooding the technological and scientific landscape (Anderson, 2008). The main thesis the text supports is that the application of techniques for massive data makes one of scientists' fundamental activities unnecessary, namely the construction of models that explain the associated reality. In Anderson's article George Box's famous maxim (*All models are wrong, but some are useful*) is confronted with Peter Norvig's *All models are wrong, and increasingly you can succeed without them*.

«77» Anderson's thesis partly justifies the data scientist's temptation to work with systems without worrying about whether there exists a (domain) theory to support that their commercially valid products are *correct*. They do not care about models because they do not really need them. Furthermore, engineers do not need an explanation of the validity of their decisions (mainly justification) because it actually does not add value to the product. It is not a mild concern; systems of this type will make (or are making) decisions that will seriously affect our rights and daily lives. The absence of models (to get scientific explanations) can cause serious defenselessness, especially if the systems are used in sensitive fields such as Predicitive Policing (Hung & Yen, 2020).

«78» Therefore -according to Anderson- we are faced with the surprising conclusion that even correlation is enough, we can forget about causality. Consequently, social, psychological oriented systems are dispensed of providing causal/scientific explanations for justifying engineering decisions (or events). For instance, we do not need to know why people behave as they do if we can measure their behavior with

accuracy and draw valid consequences from those measurements by applying mathematics (*the numbers speak for themselves*).

«79» In conclusion, in the PetaByte Age, DS teams seem condemned to refrain from building models and validating them because massive data mining would suffice for their purposes. They use ML to offer solutions such as *oracles*, implicitly solving a foundational dilemma: Would we like to know why it happens reliably or understand why it happens at the expense of losing experimental reliability? The first option allows us to act, while the second allows us to design strategies to adapt. It is clear that this dilemma has a strong impact in XAI efforts. We advance here a serious drawback to Anderson's thesis, which impacts on the Data Science practice. As Regt points out (de Regt, 2017), even in the hypothetical case of having a perfect oracle that guarantees empirical correctness, our system could be epistemologically weak. The oracle availability would not dispense the data scientist from the need to open that black box. The scientist needs to understand (apprehend a general scheme of the theory governing that oracle) to be able to *qualitatively recognize characteristic consequences of T without performing exact calculations* (de Regt, 2017).

6.6.1 Supporting Anderson's Thesis

«80» Alternative reasons can be provided to support the basis on which Anderson builds the argument. It has already been commented that in BD it is not uncommon that one of the first problems is that data scientists themselves are incapable to claim the specific hypotheses to be tested until they perform a first exploration stage (that indeed is based on data curation). This peculiarity leads us to conclude that the classical approach to explaining (validating the hypotheses obtained from models) could be inadequate. Another way to justify this would be arguing that a model is not defined because the dataset is the *digital mirror* of a CS. As the data scientist does not know how to reconstruct an image (a model) of the complete CS, he/she is merely applies ML to find interesting features of the dataset. This way is how the data scientist explains why a particular event occurs. Even if an explanatory ML technique is selected, it does not guarantee the soundness of the explanation due to previous decisions as in the curation phase (as discussed in Sect. 6.1), which outlines the starting conditions in turn.

«81» Even if the reader acknowledges the great misgivings, one should accept that Anderson's thesis is partly right in the DS practice. His statement *We don't define the conditions of the experiments, so we don't know what we're capturing* is true insofar as the exploration and the analysis in BD do not always have a starting specific goal. There's no actual knowledge to validate, but rather the reasons why the data is so, and also to infer properties of *reality* from this analysis.

«82» At this point, and limiting ourselves to XAI, one would add an (intermediate) third proposal to elucidate the dilemma of Paragraph 79. The idea is to achieve acceptability by finding the justification of the decision taken by the AI-based system in each case, the local solution to XAI (where the explanation is synonymous with outcome justification). Moreover it would be necessary to opt for an interpretation of the concept of justification as something that makes belief objectively more likely to be valid, as opposed to another interpretation of *explanation* as something

that adequately points to belief in the truth. Belief is justified by the fact that it is properly held or based on an adequate method, to the extent that truth is the objective or the norm, a *proper-aim justification* (Graham, 2010).

6.6.2 Absence of Models Versus Veracity

«83» Anderson's controversial thesis was widely contested [see e.g. Barrowman (2014)]. Indeed his arguments are weakened when one needs particular requirements. For example, when it desires data processing systems allow the data scientist to extract some solid methodology. This is something that scientific theories should do quite well, according to Popper's insights: may serve to predict like a predictor or also as a key, to tell us what would happen if some important factor is changed in our model.

«84» The absence of models in DS affects four essential scientific dimensions: the causality mentioned above, confidence in the results, the applicability of the model to data other than those used in the training phase, and finally, its ability to explain what is happening. Following our idea sketched in Paragraph 82, the explanation becomes a simple -but limited- explanation of the particular event (local XAI), even accepting the risk that it may be *defective*. However, the challenge that comes from data curation persists: how do we estimate the suitability of the dataset used (once the Extraction-Transformation-Loading process is applied) to synthesize the solution? The question to address is, in fact, the veracity of the dataset.

«85» The working under the absence of models, the data curation dependence, and the scale versus the semantics challenge suggest that the *Veracity* becomes a fundamental dimension to consider in BD besides Volume, Variety and Velocity. In some sense, Veracity is the notion built to tackle the problem of the gap between Science and Technology in massive data processing.

«86» When an engineer is working with traditional databases, he/she assumes that the domain is soundly represented by data, being the data a model. In contrast, in BD it is usual to work with *untrue* datasets: hetereogeneity and unstructured data, missing data, data distortion, incompleteness, noise, etc. These are an actual shortcomings that cause the loss of the security in the inferred results, offered by traditional databases, damaging the link between data and actual entities to that model.

«87» We adopt the meaning of Veracity as referring to how precise or *valid* a dataset would be, that is to say, to the fidelity of the data concerning the reality that they represent. However, in the context of BD, the term has another additional meaning. Veracity would also encompass the question of the reliability of the data source, and the confidence in the data processing. These are issues to be studied as they play a more relevant role in several questions: biases, anomalies, inconsistencies, and others associated with the processing itself. It becomes a critical issue to study in the new systems AA (2015), and mandatory if the agents aim to abandon the idea that ML is *data alchemy* that exempts the explainer to be accountable.

«88» It is particularly relevant the distinction between a *rough* veracity meaning, associated with the confidence in the digital picture that data represent in DS practices (quality, safety, accuracy, completeness of the information, etc.), and more formalized concepts related to the *correctness* and *validity* of the results. Related to

the latter, databases could be understood as *formal models* of the set of definition schemes that govern them [a foundational principle accepted in Deductive Databases Theory, and also in Philosophy of Science (Leonelli, 2016)]. The latter is, in turn, a formal theory that represents knowledge about the universe from where the data were extracted. If the data scientist identifies both notions, Veracity is closely related to the well-known problems of Knowledge Engineering (or even if a certain standard in database definition schemes is required, to the Semantic Web (Alonso-Jiménez et al., 2006). Since the explanation synthesized is based on the data, it depends on truthfulness in both directions.

6.7 Foundational Issues

«89» Bearing in mind the warnings about the veracity and absence (of use) of scientific models in DS discussed above, the reader may agree that solving the explaining problem may hide the problem already mentioned of confusing data with reality. There exists a common foundational issue suffered by the solutions assisted by CAIS. Data scientists do not work with the problem they intend to solve (which may be sociological in nature, for example), but with the data that the problem leaves as a *digital footprint* instead. From this source they resort to the effectory capacity of DS, and thus solutions rest framed by that. Accordingly, it is reasonable to conclude that the solution offered -the explanation shown about the phenomenon of CAIS behavior- will be intrinsically limited by that datified image, which we call the *digital shadow*. Even if agents attempt to build a scientific model using data and AI solutions, they could doubt if the source, the digital shadow, has not distorted the informational structure of the reality.

«90» A canonical example is the identification of an individual with the collection of information about him available in BD repositories. The question is not whether the identification is valid (an issue for Privacy researchers), but how identities and experience (dis)appear from BD (Ricker, 2017). The overall digital shadow of a CS is only the source from which to provide plans and actions to be applied on the CS. In the era of hyper-connectivity and ubiquity, data scientists are still chained at the bottom of the cave envisioned by Plato from which we only perceive the (digital) shadow of events or objects, ideas and concepts that move through technological reality. The shadow is made up of an unmanageable amount of data that reflect the dynamics or the *form* of these entities. It is also reflected the (social) customs and attitudes of our fellow human beings and serves to nurture AI-based systems. And as in the cave myth, it is with the shadow that we try to extract properties and *understand* the original object/event. Please note that the issue is not a sort of perspectivism. It is rather a kind of technological solipsism; the source where perspectival DS practices are nurture.

«91» Unfortunately, the risk of confusing properties from the shadow with properties of the study object will be thus persistent; it could even be considered Anderson's thesis transcript. Following the metaphor, massive data also suffer distortions as the shadows. The detailed information available, provided by massive data availability can present the data scientist with the illusion that it owns a faithful

representation of the object/event, although it is only information about it, and as such, subject to diverse constraints. On the one hand, to interpretation (particularly, to its intentionality itself). On the other hand, and related to the former, to the context from which it was extracted and to the perspectival data scientist labor. Moreover, it is constrained by the engineer's ability to properly work with the available data. Such factors would limit the capacity of the AI system to extract *knowledge* from that information, particularly when KL-inspired explanations are sought.

«92» M. Janssen and G. Kuka claim that the greatest risk is that data *become reality* (Janssen & Kuk, 2016). One could assert that a defective or incomplete treatment goes unnoticed if reality is not observed. Hence, it becomes essential to compare the results with data from reality whenever possible. Results from each step of the Data Science project should contrast with that. It is this confrontation that could require a deeper knowledge of the situation. And, at the same time, where another opportunity for scientific theories re-appears (including for Explaining).

«93» A conjecture in this regard is based on the impression -shaped by the practice- of that the closer the data are to represent the intentionality of the system/problem under study (e.g. the study of meme diffusion within a social network can be addressed as the study of a social network devoted to spread memes), the more successful the actions are. We also claim that, possibly, it is due to the human perception of epistemological similarity between the digital shadow and the ingredients the phenomenon it observes. An example would be the provisioning of AI-based powerful tools for monitoring social media opinion/sentiments that are constantly improving, and new challenges addressed (Cambria et al., 2013). Nevertheless, when the application scope is more anchored in the CS's physical reality, the shortcomings of AI-based systems are more clearly observed. They are concerns -wicked problems- as humanitarian crisis (Meier, 2015), political/sectarian violence or physical sociological phenomena (Subrahmanian & Kumar, 2017), terrorism (Johnson et al., 2015), food crises, climate change or sustainable development. It is no wonder they belong to a class of problems that need a major interdisciplinary collaboration to address many relevant aspects as disagreements about what the problem actually is, or even the existence of contradictory solutions (Rittel & Webber, 1973). This kind of complex problem requires more than the massive data processing, even challenging the interdisciplinarity itself. Due to the hype phenomenon around AI, the idea that AI-based processing suffices to solve such challenges is so widespread that stakeholders rely on (and finance) this type of application of dubious effectiveness. This is a way of falling in the absence of models issue.

«94» Yet to be discussed is whether some of these foundational issues can be tackled using weapons of similar abstract level in selected case studies. A potential option in this line would be the design of (phenomenological in nature) scientific models having inputs/outputs similar to the event/system, to both explain and provide the causal dependencies between input/output. These are truly XAI-driven surrogate models as those for modeling explaining in complex neural networks. The preventive objection would be that even being mathematical sound, a phenomenological model like this is not necessarily a purely explanatory artifact. Genuine explanatory models attempt to describe something more, namely the mechanism responsible for the various regularities in the phenomenon. The difference is well

known. For example, the explanation -utilizing equations- of certain relationships between different variables may be phenomenological in the first instance but not necessarily explanatory. Equations establish the appropriate bridge between input/output data but, in principle, do not describe the parts of any mechanism, *equations do no disclose*. Since, in phenomenological-based models, the *parts* the designers postulate have no ontic counterpart in the mechanism, one should prevent speaking about them as a *successful representational relationship*. Therefore, there is no reason to avoid speaking of mere *useful fictions* (Barberis, 2012). Recall that it had already raised in Paragraph 28 where the need for metaphors and simplifications when trying to move the explainee accessible versions of the explanation [see e.g. the visual metaphor for analysis of argument with ontologies (Borrego-Díaz & Chávez-González, 2006; Aranda-Corral & Borrego-Díaz, 2010)]. Paradoxically, this kind of *model for explanation* is not purely explanatory in XAI. They are more like *extensively appropriate* models to support another explanatory theory [an example would be (Evans et al., 2021)], a more useful theory providing more descriptive answers for a wider range of counterfactual questions (what would happen if things were different) (Barberis, 2012). Moreover, other models approaching the system behavior could be built using the program itself (for example employing data mining with Inductive Logic Programming). Other facts confirm the fickleness of these models for Explaining, particularly from the sociological point of view. For example, the adequacy of the explanation from the model may depend on whether it matches its outcomes with the explainee's expectations (Sect. 2). When it is confirmatory -there is a match- factual explanations (which might be closer to an input/output explanation) are accepted, whereas it seems that when there is a mismatch, the counterfactuals can aid, although they are necessary but not sufficient (Riveiro & Thill, 2021).

7 Bounded Rationality, Explaining and Logics

«95» It is common to be satisfied with limited explanations of events without a rigorous, plausible, inference. We accept a possibly not the best one but, for example, socially admitted or aligned with certain cognitive preferences. Among the decisions humans make, non-optimal ones abound, even others not rational. Consequently, our explanations may suffer from similar limitations, even our perception itself. For example, human performance on perceptual classification seems to approach that of an ideal observer, but economic decisions (time spent, details of perception, or inconsistent and intransitive preferences) cause its deterioration (Summerfield & Tsetsos, 2015).

«96» Extensional appropriate models, as proposed in Paragraph 94, allow deploying (logically) sound tools for reasoning. Particularly, those based on solid mathematical theories [such as Logic Programming (Evans et al., 2021)]. Despite the essential differences separating the two approaches, retrieving ideas from KRR for BR (and its application in XAI) is a very appealing approach. In particular, bridge construction between both of them becomes a promising research line. There are modeling proposals for some BR techniques as (Lipman,

1999) which are axiomatic in nature. Another example could be Glazer and Rubinstein's proposal (Glazer & Rubinstein, 2012) of a logical-oriented formulation of a persuasion model. The latter can be adapted, identifying the listener with the explainee and imposing a series of conditions that the speaker (explainer) must satisfy to be persuaded (accept the explanation).

7.1 Logics for the Inference Link in BR-Inspired Explaining

«97» So far, it has been sketched the *epistemological distance* separating three fundamental elements: the scientific theory supporting the explaining, the explainee's literacy and preferences, and the phenomena to be explained itself. In this section, we discuss how to bridge them through the *inference link* of the explanation, thinking in the representation of inferences closer to the explainee's. That is, reasoning mechanisms providing successful real-world performance, that do not need to satisfy the requirements of rational inference (Pachur & Biele, 2007; Gigerenzer & Goldstein, 1996) and that are useful for Explaining.

«98» Human inference -in a general sense- is our essential tool for designing plans and actions to cushion the effects of a CS. If explainers are also human, the question arises as to whether these dealing skills can be exploited to synthesize successful explanations; moreover, whether it is possible simulate explanations similar to those accepted by humans. The problem of simulating human explanations would cover both the search of the explanation and its acceptability. Regarding the search, mechanisms as the discovering of similarities, relations or associations, generalization, abstraction, intuition, or context-sensitivity (Duris, 2018) are involved. Likewise, the weakening of the requirements for accepting something as an explanation represents one of the human flexibility skills. The approach to its formalization from KL would thus be the first step.

«99» To guide the right choice of logic for BR, we can adhere to Gabbay and Woods' *Logic Limitation Rule* (LLR) to prevent unlimited reasoning (Gabbay & Woods, 2003):

> A logic is inappropriate for actual agents of type τ to the extent to which factors which make for agency of type τ are indiscernible in the behavior of the logic's ideal agents.

According to LLR, a logical formalism for working with realistic (bounded) agents will be inadequate if it does not induce properties that essentially distinguish the agents designed under the new paradigm from UR-based agents. For example, to accomplish LLR, it is usual to mirror cognitive limitations through syntactic restrictions on the logic that outcomes effective limitation of both expressivity and reasoning. The strategy would suffice if it actually affects the behavior of the agent by limiting the inferential process, as is the case of BR-based agents. Regardless of the logic supporting the particular BR technique, it should discuss how to read Explaining as a BR activity.

7.2 Explaining as BR Activity

«100» ELES is related to the sociotechnical realm where the CAIS is applied. This fact would be reflected within the explanandum through nomological components. For instance, through knowledge that has been considered as laws of nature of the problem. In BR, the link should be even stronger, affecting the inference link as well. Simon claimed that the first consequence of BR is that the agent's intended rationality requires constructing a simplified model of the real situation to deal with it. He/she behaves rationally concerning such a model, thus being such a behavior not even approximately optimal for the real world. Focusing on XAI, the question transcripts to whether the explanation is acceptable according the simplified model, or even or is it just *an explanation*.

«101» Original Simon's BR approach can be applied to the task of obtaining acceptable explanations. The problem to be solved would be to explain -or even convince as required in ELES- the decision/observation, working with notions as *satisfactory, sufficient* or *convincing* explanation. That is the idea of conformity, meaning that the explanation given is sufficient according to some criteria. The explanation would preserve the basic structure from the KRR point of view (explanans, inference link, explanandum) and should base on consensual knowledge by both agents (that comes from observation beside the knowledge shared by both, eg. laws of nature in the KRR sense discussed above). From the BR approach, one has to prevent the use of abstract models or AI optimization techniques where the solution is reliable if it has the whole universe. For instance, a BR-inspired explanation should (implicitly or explicitly) contain data description and the inference processes used by the explainer, since this circumscribes the context where agents worked.

«102» The explanation synthesis under BR will own specific characteristics. To study the feasibility of an ideal *provably optimal explainer agent*, it must carry out the following tasks, that come from a refinement of the theoretical framework designed by Lewis et al. (2014). Firstly, specify the environmental properties in which the explanation will be built. Secondly, design the utility function on the behaviours (which should consider factors that influence the acceptability and confidence of the explainee in the explanation itself). Thirdly, it needs to specify the type of representation and processing models that will be used. And lastly, the model must be *constructible* (according to bounded agent guidelines).

«103» Instantiating the ideas of Lewis et al. (2014), three types of theoretical scenarios can be distinguished, where the explainer may work, according to different BR constraints. *Optimality explanations* would represent those produced by the explainer with no (machine) limitations. In *Ecological-optimality explanations*, the environment where actions are decided is governed by a given distribution, but without limitations to information processing. That is, some distribution inherent to the input data is consensual between explainer and explainee, but no bounds are imposed to processing. In *Bounded-optimality explanations* there is some limitation to information processing, which reduces the repertoire of accessible solutions and the associated explanation, the *policies*. Lastly in the *Ecological-bounded-optimality explanations* both policy space and information processing are constrained.

Acceptability would depend on whether the expected behaviour resulting from the analysis corresponds to the observed behaviour.

«104» A phenomenological factor tied to environmental information accessibility should also be considered, mainly the limitation of observation and/or classification of the event itself. It has already been commented on the role of the inherent limitations of perception (Paragraph 14). If the explainer observes abundant information then shortcomings would be naturally imposed (giving rise to curation and perspectivism practices). These will also affect the efficient codification of information relevant to decision-making, which in turn affects the choice of the best action or strategy (Summerfield & Tsetsos, 2015). This would complement Lewis et al's framework, by including inherent limitations of the perception itself for Explaining.

7.2.1 Variety, BR and Ecological Rationality

«105» The psychological factors sketched in Sect. 2 already justify the need to consider human (even Human-Computer-Interaction, HCI) factors. They are not necessarily linked to the usefulness/goodness of the explanation; rather, they would be related to evaluating the utility or acceptability of BR-based explanations. One of them lies in the fact that an example set could be accepted as justification by the explainee if it offers *variety*. For instance, an example collection covering very different situations (Landes, 2020), with some completeness appearance, would enhance the explanation. Our preference for diversity -associated with the perceived completeness of the case set- can play in favor of the explanation acceptance.

«106» Nevertheless, the variety requirement would hide a balance problem. More diversity in exemplary cases requires more computational resources or more knowledge about the environment than the explainer has or can recover. This need is, in practice, opposed to an intriguing phenomenon in Ecological Rationality (ER) (Goldstein & Gigerenzer, 2002): How could more knowledge be no better -or worse- than significantly less knowledge? ER is a particular case of BR practices that contrasts with the classical notion in the social and behavioral sciences such as economics and psychology. The theory of rational choice holds that practical rationality consists of making decisions according to some fixed rules, regardless of the context. In contrast, ER asserts that rationality is essentially context-based. Studies on ER show how humans use what we know in an environment under limited resources. Also, it focuses on the match between an heuristic and the structure of the information in a particular environment. Whilst one of the priorities in Rational Choice Theory is the internal logical consistency, ER focuses on the (external) performance in the world. This aspect moves this conception further away from any notion of logical validity (in the KL sense). We could view ER as the counterpart of the theory of situated agents (Suchman, 1987) for BR, where the explaining under variety constraints would be an ER-based practice. The understanding of an expert behaviour in presence of data -an ER topic- aids to support BR-inspired explanations.

«107» The *epistemological variety* represents a psychological factor (and prospective object to study in BR) that strengthens confidence in the explained hypothesis (Landes, 2020). The variety of tests that can be considered in a BR model for Explaining can be grouped in two levels: as a variety of explanations

in a given context, and as a variety of contexts that validate the explanation. Such a variety would also be affected by the BR-based selection techniques. We do not discuss this topic here. According to Landes (2020) the solution may not be sound in general and it must be handled with care.

«108» Analogous features have to be considered for the inference link. Psychology research on heuristics in human inference processing reveals a compendium of skills for which the classic (computational) logic paradigm is not useful to explain the success of several of these, as the Recognition Heuristic (Goldstein & Gigerenzer, 2002; Todd, 2007). Hopefully, the skills could be both usable and acceptable in the explanation process. For example, the idea of applying BR techniques to *tame* the CS, has already been considered. This is done by analyzing the expert's behaviour and reflecting on the process itself. From there, the selection of attributes/characteristics in the decision making process is justified. For example, in the management system of electrical networks (e.g. in *smartgrids*), it is being considered *to imitate* the behavior of engineers in current management National Academies of Sciences, E. and Medicine (2016) (an ER activity). One of the techniques is the so-called *Principle of effective simplicity* (from BR): experts can select a relatively small number of variables and observations to diagnose, explain, predict and make decisions. An adequate modeling of the principle could speed up these activities, find useful explanations in the future, and even automate them. However, the main limitation of the explanation lies in that the explanation support is strongly human-dependent; some sort *argument from authority* due to the incorporation of expert pragmatics in explanans (Vassiliades et al., 2021).

7.2.2 Fast-and-Frugal Techniques for Explaining

«109» One of the BR challenges is the modeling of the human expertise to select one or two causes from a, sometimes infinite, number of them, to build the explanation (Miller, 2019). Similarly, explanations are selected (in a biased manner) based on the idea that people do not usually expect a complete and faithful causal explanation.

«110» The *Fast and Frugal* (FaF) methods (Gigerenzer & Goldstein, 1996) specify how the information is searched (*search rule*), when the information search ends (*stop rule*) and how the processed information is integrated into a decision (*decision rule*). These approaches soundly work due to their simplicity. Also, it provides regularity in the face of the heterogeneity of the available data. The FaF techniques produce explanations that can benefit from tools to model and evaluate the strategy followed (Phillips et al., 2017), selected from the *adaptive toolbox*, in order to design transparent assistance systems for decision-making (Raab & Gigerenzer, 2015). In Table 8 some FaF techniques are adapted to be used in Explaining. On the negative side, explanation strategies based on FaF heuristics have the risk of falling into *Cherry Picking*, *False Causality* or *Sampling Bias* fallacies, all of them related to the initial constraints imposed in FaF.

Table 8 FaF techniques that can be applied within XAI

FaF technique	Adaptation to explaining
One-Reason Decision Making Gigerenzer et al. (2008)	Explanations based on a single reason (although other can be based on another one
Recognition Heuristic Goldstein and Gigerenzer (2002)	Helps to choose between two explanations when only one of them is recognized, according to plausibility criteria (possibly a metric)
The effect less is more Goldstein and Gigerenzer (2002)	To choose the explanation that, using less information produces better results according to some measure of utility

7.2.3 Generalization by Abstraction

«111» The generalization of explanations (for their reusability) depends on the availability of more data from multiple sources, which also allows the development of richer models and greater understanding. However, when more data are available and curation is absent or defficient, models can become more complex and too detailed to be understandable by the explainee.

«112» Another risk for explaining models' reusability that comes from (BR-based) data curation is that data bias may lead to the inability to replicate studies for similar problems. This inability undertakes the explanation acceptability itself (Janssen & Kuk, 2016). A solution could be the generalization, although the level of abstraction of the explanation can condition it. Premises or conclusions, that are too abstract or general, could compromise both the actual explainee's understanding and its practical value. Abstractions may simplify explanations, but the discovery of sound abstractions is very challenging (as well as sharing its understanding) (Gunning et al., 2019). Such difficulties could lead to a greater gap between scientific rigor and practical relevance. We could claim that BR and generalizations can play opposite roles in the explanation of CAIS behavior.

7.3 Perspectivism and Curation as the Basis for BR-Based Explaining Strategies

«113» It has been emphasized that the agent's understanding of the environment is a key factor in BR approaches as ER. The selection of features and the available background knowledge on the environment leads to work within a particular context to build the explanation. What is more, the application in Explaining of BR techniques such as contextual selection or effective simplicity leads us to consider that the explanations coming from them are also perspectival in such sense. By framing a context/perspective (induced by the understanding of the phenomena), it can be assumed that the mental space where the agent's reasoning occurs would be circumscribed to that. Among other consequences, the explainers will select from the available data those that their scientific training indicates them that they are *causal*,

employing BR skills (possibly unconsciously) conditioned by the perspective. The advantage of achieving a consensual perspective lies in its status of ontological commitment about the information ecosystem, having the explanation and the results strengthened and accepted (because facilitates the internalization, Paragraph 24). However, it has already been mentioned that there exists the temptation to explain employing only statistical-computational relationships and principles (some sort of extreme empiricism) to shape the perspective (by means of estimations, bounds, thresholds, etc.). The adoption of BR practices increases the risk of the emergence of perspectives according to non explicit principles.

«114» An example of sound (techno-)perspectivism we refer to is the explanation of the so-called *Arab Spring* (years 2010–2013) by the western media as a social movement claiming social and political rights [admitting other political factors (Korotayev, 2014)]. Wikipedia presents a concrete incident in Tunisia as the spigot for the mobilizations (see https://en.wikipedia.org/wiki/Arab_Spring). One could ask whether this interpretation is not a very tight corset. The thesis is product of a perspective taken by the analysts (perhaps on political wishful thinking). There should exist a *spigot* if the system is in an unstable state, but it might not be *the cause* even if we admit it as a causal explanation. Lagi et al. (2011), through analysis of available data, founded propose another (contributing?) cause. Their analysis shows that the timing of the violent protests in North Africa and the Middle East in 2011 (as well as the earlier riots in 2008) coincides with large increases in global prices of basic foodstuffs of the most vulnerable populations. They even provide an estimate of the price threshold above which riots break out. The example clearly shows that data curation by experts is necessary, rather than massive data analysis to provide acceptable explanations, and also how these can be confronted or need to reconcile with others supported by Social Science. Also, it is an example of the perspectival application of Data Science to outcomes alternative explanations.

8 Conclusions and Future Work

«115» This work reflects the authors' conceptual journey from AI to a framework where XAI is observed as multidisciplinary in nature. We have discussed the need to adopt particular viewpoints within XAI on two problems: The XAI practices within the Data Science (and BD) universe, and the peremptory need to transmit the explanation (even the trust) on the CAIS to the stakeholder.

«116» Regarding the first of the problems, it has been pointed out the risk of a historical research and development of the XAI. The issue is worrying when DS teams cling to extreme empiricism and fall into the temptation of working without (scientific) models on the reality they study; particularly problematic when the issues they deal with are sensitive for citizenship.

«117» The second is intimately linked to the proper use of CAIS as decision-making assistant, but also as a tool for monitoring or managing CS issues. Our thesis (formed from our standpoint as researchers in KRR-based AI) claims that it is necessary to incorporate the astonishing corpus on Explaining from Philosophy of Science and Technology. The claim does not limit itself to the general principles;

it also covers its use to drive the implementation of new technology for XAI. Concerning this issue, some guidelines have been outlined for the case of exploiting BR techniques in XAI. It is interesting for ELES, which represents a socio-technical system when the explaining challenge can become a problem rooted in several issues (for instance, the stakeholder's literacy).

«118» The development of XAI is spoiled by the incipient and ongoing problems that the widespread use of CAIS is causing in society. The urgent need for explanation (which frequently hides others as that of verification, validation, or certification) means that engineers do not have time to devote the effort needed to achieve actual interdisciplinarity in XAI. The authors hope the paper can convince the AI colleagues that purely technological development can be fast but suffer real shortcomings (that affect their usefulness, safety) that are also rooted in foundational issues and not only in purely pragmatic issues.

«119» The suitability of some formal and philosophical conditions under which BR ideas can be applied in XAI have been investigated. This issue has been treated in a general way, emphasizing its philosophical, computational, and particularly AI dimensions in the field of DS and CS. We have focused on DS socio-technical systems, in contrast to other studies more focused on computational aspects of the decision itself (cf. Främling (2020)). Due to its socio-technical complexity, ELES is a paradigmatic case. The systems and the engineers could not offer an adequate explanation for stakeholders, even it may be doubtful that the technical explanations actually correspond to what happened due to perspectival principles. Moreover, it has also been analyzed the convenience of considering BR techniques to synthesize explanations that may be acceptable to the explainer, although they may suffer from deficiencies derived from BR itelf.

«120» Thus, the relationship between explainability and replicability has not been discussed in depth, even recognizing that the latter represents a good option to achieve the explanation acceptance Guttinger (2020). We could claim that the explanation of CAIS outcomes to manage CS and could mirror some of the features of the reproducibility crisis, which are becoming more common in modern Physics. However, the identification of the fields to which the replicability standard applies or not is a challenge. Guttinger (2020) argues that (at least) three different aspects of scientific practice could be used to properly answer this question: the type of questions addressed, the setup used, and the nature of the objects analyzed. From the analysis of CS and the nature of the concept of *acceptable explanation for the stakeholder*, we can conclude that XAI for working with CS seems to be framed to that grey zone of research practices where there might not be a clear answer to the replicability issue. A case-by-case analysis might be the only sensible way forward, in the same vein as Guttinger (2020). Also due to space limitations, an analysis of the status of emergence-based explanation for CS has been avoided. The techniques from Agent-Based Modeling can be combined with the macro-vision provided by proposals that exploit the epistemological nature [see e.g. our papers Aranda-Corral et al. (2013a, 2013b, 2018)]. This is a promising topic for a further research.

«121» Finally, another long-term goal to be tackled is the design of an ontology on the analytical elements that precisely define notions associated with the limitation of the agents involved in XAI. It should include concepts such as the goals,

behaviors, and different ecological and evaluation environments (Lewis et al., 2014). Understanding the explaining as a BR task, any XAI practice of this kind would be representable by specifying the elements playing a relevant role in the case of (bounded) optimal explainer agents. In this way, external agents can contextualize explanations produced in a particular socio-technical system.

«122» In addition, we think that a proper reading of critical works on Data Management practices in particular fields [e.g. Leonelli (2016)] can provide very useful ideas for understanding the data curation process. Lastly, there exists also the possibility of implementing some of these ideas into the software for XAI in Data Science. By asking what can be learnt from these practices in data science, one could extract those which overcome the epistemic losses that data curation can cause.

Acknowledgements This work is supported by Spanish State Investigation Agency (Agencia Estatal de Investigación), Project PID2019-109152GB-I00/AEI/10.13039/501100011033. We are very grateful to the reviewers for their suggestions, and for guidance on additional references that have enriched our work.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

AA, V. (2015). *The Field Guide to Data Science* (2nd ed.). Booz Allen Hamilton.

Addis, T. (2014). *Natural and artificial reasoning—an exploration of modelling human thinking. Advanced information and knowledge processing*. Springer.

Alonso-Jiménez, J. A., Borrego-Daz, J., Chávez-González, A. M., & Martín-Mateos, F. J. (2006). Foundational challenges in automated semantic web data and ontology cleaning. *IEEE Intelligent Systems, 21*(1), 42–52.

Alrøe, H. F., & Noe, E. (2014). Second-order science of interdisciplinary research: A polyocular framework for wicked problems. *Constructivist Foundations, 10*(1), 65–76.

Anderson, C. (2008). The petabyte age: Because more isn't just more—more is different. Retrieved from http://www.wired.com/2008/06/pb-intro/.

Anderson, J. D., & Perona, P. (2014). Toward a science of computational ethology. *Neuron, 84*(1), 18–31.

Aranda-Corral, G. A. & Borrego-Díaz, J. (2010). Mereotopological analysis of formal concepts in security ontologies. In Herrero, Á., Corchado, E., Redondo, C., & Alonso, Á (Eds.), *Computational Intelligence in Security for Information Systems 2010—Proceedings of the 3rd International Conference on Computational Intelligence in Security for Information Systems (CISIS'10), León, Spain, November 11–12, 2010*, Vol. 85 of *Advances in Intelligent and Soft Computing* (pp. 33–40). Springer.

Aranda-Corral, G. A., Borrego-Díaz, J., & Galán-Páez, J. (2013a). Qualitative reasoning on complex systems from observations. In *Hybrid Artificial Intelligent Systems* (pp. 202–211). Springer .

Aranda-Corral, G. A., Borrego-Díaz, J., & Giráldez-Cru, J. (2013b). Agent-mediated shared conceptualizations in tagging services. *Multimedia Tools Applications, 65*(1), 5–28.

Aranda-Corral, G. A., Borrego-Díaz, J., & Galán-Páez, J. (2018). Synthetizing qualitative (logical) patterns for pedestrian simulation from data. In Bi, Y., Kapoor, S., & Bhatia, R., (Eds.), *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016* (pp. 243–260). Springer.

Araujo, T., Helberger, N., Kruikemeier, S., & Vreese, C. H. D. (forthcoming). In AI we trust? perceptions about automated decision-making by artificial intelligence. *AI and Society* 1–13.

Aronson, D. R. (2011). *The illusory validity of subjective technical analysis, chapter 2* (pp. 33–101). Wiley.

Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of ACM, 63*(3), 48–55.

Barberis, S. D. (2012). Un análisis crítico de la concepción mecanicista de la explicación. *Revista Latinoamericana de Filosofia, 38*(2), 233–265.

Barrowman, N. (2014). Correlation, causation, and confusion. *The New Atlantis, 1*(43), 23–44.

Bas, C. V. F. (1980). *The Scientific Image*. Oxford University Press.

Biewald, L. (2016). *The machine learning problem of the next decade*. Retrieved from https://www.computerworld.com/article/3023708/the-machine-learning-problem-of-the-next-decade.html.

Booth, S., Muise, C., & Shah, J. (2019). Evaluating the interpretability of the knowledge compilation map: Communicating logical statements effectively. In Kraus, S., (Eds.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019* (pp. 5801–5807).

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). Self-driving cars and engineering ethics: The need for a system level analysis. *Science and Engineering Ethics, 25*(2), 383–398.

Borrego-Díaz, J., & Chávez-González, A. M. (2006). Visual ontology cleaning: Cognitive principles and applicability. Lecture Notes in Computer Science. In Y. Sure & J. Domingue (Eds.), *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, June 11–14, 2006, Proceedings* (Vol. 4011, pp. 317–331). Springer.

Borrego-Díaz, J., & Páez, J. G. (2022). Knowledge representation for explainable artificial intelligence. *Complex & Intelligent Systems* 1–23.

Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of autonomous systems. *IEEE Intelligent Systems, 28*(3), 54–61.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*(2), 15–21.

Craver, C. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience. Oxford scholarship online: Philosophy module*. Clarendon Press.

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*(3), 355–376.

Cugueró-Escofet, N., & Rosanas-Martí, J. (2019). Trust under bounded rationality: Competence, value systems, unselfishness and the development of virtue. *Intangible Capital, 15,* 1–21.

Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfield relations, and anomaly resolution*. Cambridge Studies in Philosophy and Biology. Cambridge University Press.

DARPA. (2016). *Explainable Artificial Intelligence (XAI) Program*. Defense Advanced Research Projects Agency: Technical report.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? *AI Magazine, 14*(1), 17.

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making. *AI Society, 35*(4), 917–926.

de Regt, H. (2017). *Understanding Scientific Understanding*. Oxford Studies in Philosophy of Science. Oxford University Press.

Dick, S. (2015). Of models and machines: Implementing bounded rationality. *Isis, 106*(3), 623–634.

Díez, J. (2014). Scientific w-explanation as ampliative, specialized embedding: A neo-hempelian account. *Erkenntnis, 79*(S8), 1413–1443.

Dimitrijević, D. R. (2019). Causal closure of the physical, mental causation, and physics. *European Journal for Philosophy of Science, 10*(1), 1.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. In T. R. Besold, & O. Kutz, (Eds.), *Proc. First Int. Workshop on Comprehensibility and Explanation in AI and ML*, Volume 2071 of *CEUR Workshop Proceedings* (pp. 1–8). CEUR-WS.org.

Dudai, Y., & Evers, K. (2014). To simulate or not to simulate: What are the questions? *Neuron, 84*(2), 254–261.

Duris, F. (2018). Arguments for the effectiveness of human problem solving. *Biologically Inspired Cognitive Architectures, 24,* 31–34.

Evans, R., Bošnjak, M., Buesing, L., Ellis, K., Pfau, D., Kohli, P., & Sergot, M. (2021). Making sense of raw input. *Artificial Intelligence, 299,* 103521.

Fellous, J.-M., Sapiro, G., Rossi, A., Mayberg, H., & Ferrante, M. (2019). Explainable artificial intelligence for neuroscience: Behavioral neurostimulation. *Frontiers in Neuroscience, 13,* 1346.

Findl, J., & Suárez, J. (2021). Descriptive understanding and prediction in Covid-19 modelling. *History and Philosophy of the Life Sciences, 43*(4), 1–31.

Forrester, A. I. J., Sobester, A., & Keane, A. J. (2008). *Engineering design via surrogate modelling—a practical guide*. Wiley.

Främling, K. (2020). Decision theory meets explainable AI. In D. Calvaresi, A. Najjar, M. Winikoff, & K. Främling (Eds.), *Explainable, transparent autonomous agents and multi-agent systems* (pp. 57–74). Springer.

Füllsack, M. (2014). The circular conditions of second-order science sporadically illustrated with agent-based experiments at the roots of observation. *Constructivist Foundations, 10*(1), 46–54.

Gabbay, D. M., & Woods, J. (2003). Chapter 3—logic as a description of a logical agent. In D. M. Gabbay & J. Woods (Eds.), *Agenda Relevance, Volume 1 of A Practical Logic of Cognitive Systems* (pp. 41–68). Elsevier.

Gajdoš, P., & Snášel, V. (2014). A new FCA algorithm enabling analyzing of complex and dynamic data sets. *Soft Computing, 18*(4), 683–694.

Gerbaudo, P. (2020). From data analytics to data hermeneutics. Online political discussions, digital methods and the continuing rel- evance of interpretative approaches. *Digital Culture & Society, 2*(2), 95–112.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650–669.

Gigerenzer, G., Martignon, L., Hoffrage, U., Rieskamp, J., Czerlinski, J., & Goldstein, D. G. (2008). *One-reason decision making, Chapter 108*, (Vol. 1, pp. 1004–1017). Elsevier.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT Press.

Giráldez-Cru, J., & Levy, J. (2016). Generating SAT instances with community structure. *Artificial Intelligence, 238,* 119–134.

Glazer, J., & Rubinstein, A. (2012). A model of persuasion with boundedly rational agents. *Journal of Political Economy, 120*(6), 1057–1082.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., et al. (2018). Explainable AI: The new 42? In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine learning and knowledge extraction* (pp. 295–303). Springer.

Goldstein, D., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75–90.

Graham, P. J. (2010). Theorizing justification. In *Knowledge and skepticism* (pp. 45–71). MIT Press.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai-explainable artificial intelligence. *Science Robotics, 4*(37), 7120.

Guttinger, S. (2020). The limits of replicability. *European Journal for Philosophy of Science, 10*(2), 10.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology, 36*(1), 49–67.

Hempel, C. (1970). *Aspects of scientific explanation: And other essays in the philosophy of science*. Number v. 2 in Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science. Free Press.

Hernandez, J., & Ortega, R. (2019). Bounded rationality in decision-making. *MOJ Research Review, 2*(1), 1–8.

Hinsen, K. (2014). Computational science: Shifting the focus from tools to models. *F1000Research, 3*(101), 1–15.

Hofman, J., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., et al. (2021). Integrating explanation and prediction in computational social science. *Nature, 595*(7866), 181–188.

Huneman, P. (2018). Outlines of a theory of structural explanations. *Philosophical Studies, 175*(3), 665–702.

Hung, T. & Yen, C. (2020). On the person-based predictive policing of AI. *Ethics and Information Technology*.

Hvoreckỳ, J., Šimúth, J., & Lichardus, B. (2013). Managing rational and not-fully-rational knowledge. *Acta Polytechnica Hungarica, 10*(2), 121–132.

Ihde, D. (2010). *Heidegger's technologies: Postphenomenological perspectives*. Fordham University Press.

Janssen, M., Hartog, M., Matheus, R., Ding, A. Y., & Kuk, G. (2021). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review*, 0894439320980118.

Janssen, M., & Kuk, G. (2016). Big and open linked data (bold) in research, policy, and practice. *Journal of Organizational Computing and Electronic Commerce, 26*(1–2), 3–13.

Jarke, J., & Macgilchrist, F. (2021). Dashboard stories: How narratives told by predictive analytics reconfigure roles, risk and sociality in education. *Big Data & Society, 8*(1), 20539517211025560.

Johnson, N. F., Restrepo, E. M., & Johnson, D. E. (2015). *Modeling human conflict and terrorism across geographic scales, Chapter 10* (pp. 209–233). Springer.

Jones, A. J., Artikis, A., & Pitt, J. (2013). The design of intelligent socio-technical systems. *Artificial Intelligence Review, 39*(1), 5–20.

Kim, J. (2005). *Physicalism, or something near enough*. Princeton University Press.

King, M. (2020). Explanations and candidate explanations in physics. *European Journal for Philosophy of Science, 10*(1), 7.

Klein, G., Moon, B., & Hoffman, R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems, 21,* 88–92.

Kliegr, T., Bahník, Štěpán, & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence, 295,* 103458.

Koehler, D. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110,* 499–519.

Korotayev, A. (2014). The Arab spring: A quantitative analysis. *Arab Studies Quarterly, 36,* 149–169.

Kridalukmana, R., Lu, H. Y., & Naderpour, M. (2020). A supportive situation awareness model for human-autonomy teaming in collaborative driving. *Theoretical Issues in Ergonomics Science*, 1–26.

Kroes, P., Franssen, M., Poel, I., & Ottens, M. (2006). Treating socio-technical systems as engineering systems: Some conceptual problems. *Systems Research and Behavioral Science, 23,* 803–814.

Kroes, P., & Verbeek, P. (2014). *The moral status of technical artefacts. Philosophy of Engineering and Technology*. Springer.

Lagi, M., Bertrand, K. Z., & By, Y. (2011). The food crises and political instability in North Africa and the middle east. *SSRN, 20*(1), 1–15.

Landes, J. (2020). Variety of evidence and the elimination of hypotheses. *European Journal for Philosophy of Science, 10*(2), 12.

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press.

Lewis, R. L., Howes, A. D., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science, 6*(2), 279–311.

Li, J., & Liu, H. (2017). Challenges of feature selection for big data analytics. *IEEE Intelligent Systems, 32*(2), 9–15.

Lieto, A., Lebiere, C., & Oltramari, A. (2018). The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research, 48,* 39–55.

Lipman, B. L. (1999). Decision theory without logical omniscience: Toward an axiomatic framework for bounded rationality. *The Review of Economic Studies, 66*(2), 339–361.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*(3), 232–257.

Longino, H. E. (2006). *Theoretical pluralism and the scientific study of behavior, Chapter 6* (Vol. 19, pp. 102–131). University of Minnesota Press, ned—new edition edition.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (pp. 4768–4777). Curran Associates Inc.

Margolis, J. (1983). The logic and structures of fictional narrative. *Philosophy and Literature, 7*(2), 162–181.

Meier, P. (2015). *Digital humanitarians: How big data is changing the face of humanitarian response*. CRC Press Inc.

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

Moreira, C. (2019). *Unifying decision-making: A review on evolutionary theories on rationality and cognitive biases, , Chapter 19* (pp. 235–248). Springer.

Müller, K. H., & Riegler, A. (2014). Second-order science: A vast and largely unexplored science frontier. *Constructivist Foundations, 10*(1), 7–15.

National Academies of Sciences, E. and Medicine. (2016). In *Refining the Concept of Scientific Inference When Working with Big Data: Proceedings of a Workshop—in Brief*. The National Academies Press.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*(1), 87–127.

Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE, 104*(1), 11–33.

Pachur, T., & Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. *Acta Psychologica, 125*(1), 99–116.

Páez, A. (2009). Artificial explanations: The epistemological interpretation of explanation in AI. *Synthese, 170*(1), 131–146.

Papineau, D. (2001). *The rise of physicalism, Chapter 1* (pp. 3–36).

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*(none), 96–146.

Phillips, N., Neth, H., Woike, J., & Gaissmaier, W. (2017). Fftrees : A toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgment and Decision Making, 12,* 344–368.

Pirolli, P. & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis* (pp. 2–4).

Price, M., Walker, S., & Wiley, W. (2018). The machine beneath: Implications of artificial intelligence in strategic decision making. *PRISM, 7*(4), 92–105.

Raab, M., & Gigerenzer, G. (2015). The power of simplicity: A fast-and-frugal heuristics approach to performance science. *Frontiers in Psychology, 6,* 1672.

Rago, A., Cocarascu, O., Bechlivanidis, C., Lagnado, D., & Toni, F. (2021). Argumentative explanations for interactive recommendations. *Artificial Intelligence, 296,* 103506.

Reutlinger, A. (2014). Why is there universal macrobehavior? Renormalization group explanation as noncausal explanation. *Philosophy of Science, 81*(5), 1157–1170.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (pp. 1135–1144). Association for Computing Machinery.

Ricker, B. (2017). *Reflexivity, positionality and rigor in the context of big data research, Chapter 4* (pp. 96–118). University of Iowa Press.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155–169.

Riveiro, M., & Thill, S. (2021). "that's (not) the output i expected!'' on the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence, 298,* 103507.

Romele, A., Severo, M., & Furia, P. (2020). Digital hermeneutics: From interpreting with machines to interpretational machines. *AI and Society*, 1–14.

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A modern approach* (2nd ed.). Pearson Education.

Russell, S. J., & Subramanian, D. (1995). Provably bounded-optimal agents. *The Journal of Artificial Intelligence Research, 2*(1), 575–609.

Salmon, W., & Press, P. U. (1984). *Scientific explanation and the causal structure of the world. LPE Limited Paperback Editions*. Princeton University Press.

Schupbach, J. N. (2019). Conjunctive explanations and inference to the best explanation. *Teorema: Revista Internacional de Filosofía, 38*(3), 143–162.

Simon, H. (1957a). A behavioural model of rational choice. In H. Simon (Ed.), *Models of man: Social and rational; mathematical essays on rational human behavior in a social setting* (pp. 241–260). Wiley.

Simon, H. A. (1957b). *Models of Man: Social and rational: Mathematical essays on rational human behavior in a social setting*. Garland Publishing, Incorporated: Continuity in Administrative Science. Ancestral Books in the Management of Organizations.

Slack, D., Hilgard, S., Singh, S., & Lakkaraju, H. (2021). Feature attributions and counterfactual explanations can be manipulated. *CoRR*.

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access, 9,* 11974–12001.
Stern, L. (2005). *Interpretive reasoning*. Cornell University Press.
Subrahmanian, V. S., & Kumar, S. (2017). Predicting human behavior: The next frontiers. *Science, 355*(6324), 489–489.
Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.
Sullivan, E. (2019). Universality caused: The case of renormalization group explanation. *European Journal for Philosophy of Science, 9*(3), 36.
Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences, 19*(1), 27–34.
Todd, P. M. (2007). How much information do we need? *The European Journal of Operational Research, 177*(3), 1317–1332.
Townsend, J., Chaton, T., & Monteiro, J. M. (2019). Extracting relational explanations from deep neural networks: a survey from a neural-symbolic perspective. *IEEE Transactions on Neural Networks and Learning Systems* (pp. 1–15).
van de Poel, I. (2020). Embedding values in Artificial Intelligence (AI) systems. *Minds and Machines*.
van der Waa, J., Nieuwburg, E., Cremers, A. H. M., & Neerincx, M. A. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291,* 103404.
Vassiliades, A., Bassiliades, N., & Patkos, T. (2021). Argumentation and explainable artificial intelligence: A survey. *The Knowledge Engineering Review, 36,* e5.
Wang, X., Wang, D., Xu, C., He, X., Cao, Y., & Chua, T. (2019). Explainable reasoning over knowledge graphs for recommendation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* (pp. 5329–5336). AAAI Press.
Weld, D. S. & Bansal, G. (2018). Intelligible artificial intelligence. *CoRR*.
Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM, 62*(6), 70–79.
Wong, J. C. (2020). Computational causal inference.
Woodward, J. (2019). Scientific explanation. In Zalta, E. N., (Eds.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition.

Authors and Affiliations

Joaquín Borrego-Díaz[1] · Juan Galán-Páez[1]

Juan Galán-Páez
juangalan@us.es

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S. Ingeniería Informática – Universidad de Sevilla, Avda. Reina Mercedes s.n., 41013 Sevilla, Spain