Complex & Intelligent Systems (2022) 8:1579–1601
https://doi.org/10.1007/s40747-021-00613-5

ORIGINAL ARTICLE

Knowledge representation for explainable artificial intelligence

Modeling foundations from complex systems

Joaquín Borrego-Díaz[1] · **Juan Galán Páez**[1,2]

Received: 10 December 2020 / Accepted: 3 December 2021 / Published online: 4 January 2022

Abstract
Alongside the particular need to explain the behavior of black box artificial intelligence (AI) systems, there is a general need to explain the behavior of any type of AI-based system (the explainable AI, XAI) or complex system that integrates this type of technology, due to the importance of its economic, political or industrial rights impact. The unstoppable development of AI-based applications in sensitive areas has led to what could be seen, from a formal and philosophical point of view, as some sort of crisis in the foundations, for which it is necessary both to provide models of the fundamentals of explainability as well as to discuss the advantages and disadvantages of different proposals. The need for foundations is also linked to the permanent challenge that the notion of explainability represents in Philosophy of Science. The paper aims to elaborate a general theoretical framework to discuss foundational characteristics of explaining, as well as how solutions (events) would be justified (explained). The approach, epistemological in nature, is based on the phenomenological-based approach to complex systems reconstruction (which encompasses complex AI-based systems). The formalized perspective is close to ideas from argumentation and induction (as learning). The soundness and limitations of the approach are addressed from Knowledge representation and reasoning paradigm and, in particular, from Computational Logic point of view. With regard to the latter, the proposal is intertwined with several related notions of explanation coming from the Philosophy of Science.

Keywords Complex systems · Explainable artificial intelligence · Epistemological modeling · Formal concept analysis

Introduction

Alongside the particular need to explain the behavior of black box Artificial Intelligence (AI) systems, there is a general need to explain the behavior of any type of AI-based system (the explainable AI, XAI) or complex system, that integrate this type of technology, due to the importance of its economic, political or industrial rights impact. In either case, the AI component might not be isolated, but possibly part of a broader treatment of information, or integrated into a broader Complex System (CS). The integration complicates the task of explaining systems behavior, which can be so complex that traditional systems theory thinking becomes insufficient. In areas where it is important (or even critical) to monitor the behavior of a system that includes AI-based modules, such as systems for Big Data (BD), Internet of Things, or Cloud Computing, some kind of specification of the behavior of such module or at least an explanation of the decisions taken will be needed. Please note that the need is not generalized, not every automatic or even AI-based system should be controlled. Still, there exist cases (for example when people's rights are affected) in which the system must be subject to some form of certification, traceability, and assessment of both applicability and performance. Also, the XAI itself aims for the achievement of many goals different in nature (cf. [1]).

It is clear that not every system that makes an automated decision is AI, nor every AI is Machine Learning (ML), nor everything that is announced as AI is actually AI-based there is an evident hype on the subject and the term is often used as a marketing resource or for justifying business strategies. This hype around certain techniques does not refute the

✉ Joaquín Borrego-Díaz
jborrego@us.es

Juan Galán Páez
juangalan@us.es

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S. Ingeniería Informática, Universidad de Sevilla, Spain

[2] Datrik Intelligence S.A., Sevilla, Spain

above-mentioned fact of the current ubiquity and power of AI-based systems, which makes it very common for them to be used to improve other systems. The latter inherits potential explainability problems from the former, in any of the three notions (levels) of Explainable AI (XAI) that actually exist, according Doran et al. [2]: opaque systems that offer no insight into its algorithmic mechanisms; interpretable systems where users can mathematically analyze its algorithmic mechanisms; and comprehensible systems that emit symbols enabling user-driven explanations of how a conclusion is reached.

The challenges that comes with Explainability (and its impact or relationship on diagnosis or debugging—verification—for example) are similar (of course, bridging the gap) to a crisis of foundations proper to the dizzying advance of a branch of knowledge (as occurred with the emergence of Set Theory in the last century, for example). A crisis that affects the validity of the results is usually confronted by going back to the analysis of the capacity of fundamental models of formalization and their properties. However, even sharing similarities with older challenges in Mathematics, there are some characteristics of XAI that make the problem somewhat transversal in nature. For example, its (social, psychological) inter-agent nature and its relationship with system correction.

Two natural strategies used to address the problem of XAI namely, using interpretable models or, if it is not possible, with post hoc explaining, by argumenting/explaining the result once it is obtained—may be insufficient. As argued by Miller [3], the creation of explainable intelligent systems requires addressing major issues. Firstly, Explainability can be a question interactive in nature between humans and the (automated or semi-automated) AI system; and secondly, it is peremptory the design of representations that support the articulation of the explanations is required. Tuning up more, Weld and Bansal [4] require a good explanation to be simple, easy to understand, and faithful (accurate), conveying the true cause of the event. Therefore a balancing problem between two demands (Miller's versus Weld and Bansal's ones) is faced.

The Knowledge Representation and Reasoning paradigm faces other issues that increase the complexity of the challenge. Mainly that one comes from the difficulty of translating some sort of logical explanation into a language that is acceptable and intelligible by the non-expert. In fact, we could consider that two elements need to be translated frequently, for example, when it is necessary to justify the decision taken (i.e. a complete argument). Of course the conclusion, but also the part of the Knowledge Base (KB) that has been used to entail it, that is, the initial hypotheses beside the inference links. As for the justification process itself, it should also be translated or adapted when it is not legible for the explainee (cf. [5]).

Working in a massive data framework can exacerbate the problem of Explaining. It implies dealing with problems with thousands of features (among other issues [6]), thus performance requirements are likely to force the adoption of methods that are difficult or impossible to explain in particular scenarios such as deep neural networks or enhanced decision forests [7]. It is often the case that post hoc explanations of events may be the only way to facilitate human understanding; such kind of explanations could be more easily accepted if some common human reasoning patterns, which are not direct variants of purely logical reasoning, are selected.

The use of surrogate models[1] could simplify certain aspects to make possible to explain the event/result to a wider audience as well as to discuss the reasons why concrete results are predicted or occur. The explanation could even go as far as using economic arguments such as the cost-effectiveness of the decision, to estimate the economic cost of each possibility (true/false or positive/negative outputs).

Naturally, it is not the only option. For example, other approach to research is based on treating AI instruments as biological entities, studying them from that point of view. For example, considering neural networks as experimental objects in biology, rather than as analytical and purely mathematical objects (see e.g. Bornstein's [8]). The approach would involve, for example, analyzing the individual components, disturbing some inputs or sectioning parts in order to check their role and the elasticity of the model according to the ideas from Neuroscience.

The construction of models to support explanation—especially for ML-based intelligent systems—is a challenge that embraces several techniques ranging from logical causal models (as the already mentioned tradition of classical Expert Systems) to those specialized in deep learning (cf. [9]). To approach the issue from Knowledge Level (KL), the need to reconcile two levels of (representation and) reasoning (for the explainer and the explainee) through some kind of accepted, consensual models, becomes more pressing. What can Newell's KL [10] paradigm offers to meet the challenge of explainability? Mainly explanation, interpretation, and justification, which are activities deeply rooted in AI research, as they provide reliability in systems with autonomy in the decision-making process. At KL, a first source of explanation comes from KL's paradigm itself.

The idea of surrogation can bridge several levels of sophistication between predictions and experiments. By helping to understand the event or system, the surrogate model helps to interpret or justify decision making by drawing from data source. In the case at hand, the gap would be epistemological in nature. A KL-based surrogate model should bridge the

[1] This term is used to denote the general idea of simulating the behavior of a given model.

response offered by the AI-based system and those that are acceptable to humans. For example, expert systems can be considered as surrogate models for human expertise. As well as trying to reproduce the expert's results, they also map and make explicit both the knowledge used and acquired.

In KL paradigm, agents/systems work mainly with logical formulas or symbolic expressions that seek to represent information from the world, to obtain conclusions about it, by mechanized symbolic manipulations without any intended meaning. All that is needed is to specify what the agent knows or believes and what its goals are. Therefore, by considering the idea of KL-based surrogate models, the separation between the logical abstraction and the implementation details (including the implementation of the inference/decision process itself) has to be assumed. Newell's proposal for KL was intended to clarify the elements that should be considered in order to formalize the idea of rational agent by separating the two modules for the purpose of studying without ambiguity two problems: Knowledge Representation and Reasoning (KRR) problems. Davis et al. [11] explicit as one of the roles of the Knowledge Representation that of surrogate, which serves as a substitute for the original, to reason about the world and infer the decision to be made. Of course it is not the only one. It is also useful to represent ontological commitments (including background knowledge) and serves as theories for reasoning. And the one that will interest us for the paper: an environment in which information can be organized and where agents can think. More elaborate theories on knowledge taxonomy such as Addis' [12] (p. 46) further break down these roles.

Classic surrogate models are useful for the expert. However, their adaptation to face the problem of explainability can raise questions with differentiating nuances. On the one hand, the qualitative nature of human reasoning, its cognitive and functional limitations, make rational conceptual and qualitative explanations easier to follow and accept by non-specialists (clients/users, legislators, planners, product validation managers, etc.). Therefore, any numerical solution should be adapted to one of those if possible. On the other hand, this type of explanation is approximate, and tends to sacrifice rigor for the benefit of its understanding. It might be difficult to find a balance between the robustness of quantitative models and the local, qualitative, approximate, or even example-based ones that can be used in the KRR paradigm.

Focusing on the model that supports our explanation, similar questions to other situations where other kinds of surrogate models are considered arise [13]. Among those of interest for the paper, and thinking for explaining: What data and sample are used for the explanation provided by the surrogate model? What approximation method should be used? How is the surrogate model that produces the explanation? What if there's a discrepancy between two different explanations? The soundness of a surrogate model designed to bring the theoretical model closer to an explainable one will, therefore, depend on an adequate response to each of the questions.

The considerations set out so far are intended to outline the challenge of formalization that XAI represents. From a broad perspective, factors of very different nature influence the treatment of explainability and make clear the need for foundational (and also interdisciplinary) analyses. A desirable objective would be to accommodate formal and philosophical discussions and proposals in the same model. Although it is utopian in its generality, any model that serves as a partial solution would be welcomed by the scientific community.

Motivation and aim of the paper

As it can be guessed from the introduction above, the paper discusses some foundational ideas from Knowledge Representation in AI for explainability, as well as how these could be studied from a formal point of view, despite their various edges and implications. The paper is devoted to the design of an epistemological model to support some of the notions discussed in the first part. The model is the basis for the different activities that epistemology applied to CS states: from the representation of systems from a qualitative point of view to the simulation and subsequent prediction about their future. The aim is to show that the model can be an useful tool for representing both the theoretical concepts involved in the process of explainability and for addressing associated problems. Ideas toward a formal epistemological model for foundations of explaining are discussed, as well as how related issues are represented in such a model, and philosophical considerations. We believe that it is necessary, to soundly outline a general vision of the document, to briefly review this model, which will be addressed in detail in the second part of the document.

Towards an abstract epistemological model

As stated above, the main motivation of the paper is to establish a common formal framework to serve as theoretical model that can be used to discuss foundational (formal, philosophical) issues in Explaining. Starting from a purely perceptual scenario, a model for specifying several notions is built. Due to the heterogeneity of the different types of CS (which include some of the more sophisticated AI-based systems), there are no general mechanisms for addressing issues related to systems that may be essentially different in nature and purpose. However, if an abstraction with the language of events and observations is made, it is possible to provide a mathematical model that allows illustrating, in a logical-mathematical language, some foundational properties. Partial versions of the framework have been presented by authors and applied in several case studies of a different

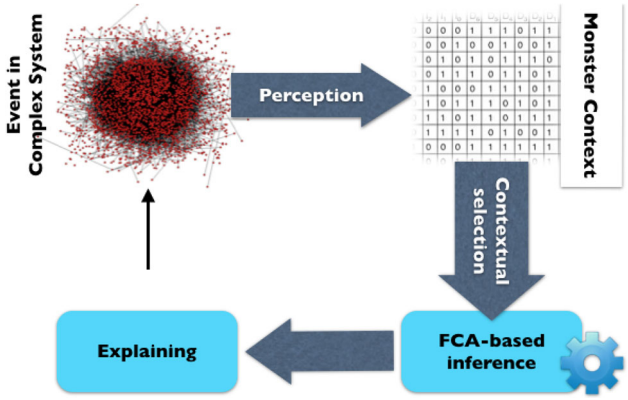

Fig. 1 FCA-based model for (qualitative) reasoning with Complex Systems

nature [14–18]. However, until now we had not presented a unified vision of the model, as well as its contextualization within the problem of explanation. Its theoretical character aims to clarify some of the ideas and issues discussed in the paper, thus providing a framework and/or a set of guidelines to be followed for the modeling and qualitative analysis of CS and related problems. In Fig. 1, the proposal (which comes from the philosophical and computational study of CS) is depicted. For now, it is enough to say that the idea is to mathematize the universe of data that comes from perception and the need to contextualize the part of that information that will be transformed into 'knowledge' to reason about the system.
Step 1: Extracting and fusing data

The model starts from raw data (right arrow in the top, Fig. 1), understanding this as everything that is perceived and analyzed, as raw information (at the bit level, if desired). In other words, information is accumulated with the maximum possible granularity and then fused into knowledge. This way we do not presuppose a conceptual representation and reasoning language beyond the one that comes from direct perception since, as argued, prefixing it would limit the vision of the events. Some questions arise immediately, as Why avoiding any high-level specification in the first instance? Why resorting to bit-level event observation (including input–output data) to indicate their basic properties? The answer is that it is not desired to limit the model by the selection of a language, beyond one denoting the very basic attributes of perceived signals. Also, we intend to offer a universal model for justification of notions and thus it is not to presuppose features inherited from a previous choice of language. The idea is that conceptual language emerges from the analysis and selection of characteristics, and its description (and further explanation) should be easy to accept by the observer.
Step 2: Contextual selection

Selecting the right context in which to find the explanation (fusing the associated information) is the basis that would serve to build perspectives using an algorithmic combination of the raw information. With this aim, Formal Concept Analysis (FCA) [19] has been chosen as a building tool (box, downright in Fig. 1). It will be used, for example, as an interface between the perceptions and the system that generates the response (explanation in our case), building what one might call event/observational concepts. Some concepts are spontaneously activated or made available for use solely based on a subject's being in a certain perceptual state [20]. To avoid the so-called content inflation [20], the analysis should be delimited, thus a sub-context is selected from the global context of all perceptions (right down arrow, Fig. 1). In this way, the explaining synthesis is confined to what the agent believes or decides that plays some role in the explanation.
Step 3: Explaining: Machine Learning tools ideas are applied on the contextual selection and using FCA-based semantics.

Please note that the paper should not be considered a research work on FCA. Such theory is used as tool to formalize some of the notions through an abstraction of the epistemological reconstruction of CS from available observations. This type of approach allows the simulation, validation and prediction tasks to be carried out based on the model. Furthermore, justifications based on logical mechanisms with relatively less complex representations, enabling the use of traditional KRR techniques such as Inductive Logic Programming (ILP), will be presented. In short, a theoretical model is proposed, with a foundational vision, to design a surrogate meta-model based on the KL paradigm that allows formalizing and discussing some of the (philosophical and formal) notions about observability, justification, argumentation and explanation that we will be dealing with in the paper. It could be considered general methods that allow us evaluating the quality of the models (both their structure and the information they contain).

Structure of the paper

After the preceding introduction, the paper starts noting some observations about the role Bounded Rationality (BR) can play in solving the explanation problem. Thus, the factors that influence what the explanation under BR would be and how the explanation techniques could be applied in that case are analyzed, including the perspectivist nature of explanation inherited from BR's role in the process, and their relationship with the so-called data curation (see "Perspectivism as explaining strategy"). "Towards a theoretical general model for explaining phenomena: background" is dedicated to motivating the choice of FCA as a tool free of specific semantic ties to analyze the problem, and to make the paper largely self-contained—the basic elements of FCA are introduced. "A toolbox for specifying explanations in complex systems" represents the formalization of the model and the introduction of a basic explanation format inspired by ideas from

the Argumentation. From this notion, and combining it with others of Inductive Logic Programming, the different elements are represented. The foundational nature of the paper is highlighted through the analysis of the model's behavior in the face of the extension of available information. The paper ends with some considerations about the proposal and future direction of the extension of the work.

Some notes on bounded rationality and explaining

Contextual selection (and possibly subsequent reasoning) will be guided by techniques that reduce the search space and decrease the complexity of reasoning. For example, those based on bounded rationality [21] may be useful. Although it is not the aim of this paper to break down such techniques (nor to specialize them to our specific model), we do believe it is interesting to comment on certain issues in this respect.

Following the analysis carried out in Lewis et al.'s paper [22], three types of explanations that could be produced by an explainer agent can be distinguished: Optimality explanations (no machine bounds), Ecological-optimality explanations (the environment where actions are decided upon responding to a given distribution but there are no limitations to the processing of the information), Bounded-optimality explanations (limitation to information processing, which reduces the repertoire of accessible solutions and the associated explanation, the policies), and lastly the Ecological-bounded-optimality explanations (in which both policy space and information processing are constrained). Therefore, to the extent that the expected behavior or the structure of the policy resulting from the analysis corresponds to the observed behavior, then the behavior has been explained in each context of explainability.

Formal epistemology, complex systems and BR

Data collection and processing are key daily tasks in CS with the aim of to obtain a reasonably accurate and concise approximation of the system and its behavior (that could lead us to a surrogate model), so that we can understand it to some extent [23]. If one wishes to explain the events that perceives, it is natural to consider an approach similar to formal epistemology. Moreover, if one wants to extract actionable knowledge, the natural approach would be the Applied Epistemology that KL would represent.

It has already justified that the use of techniques and ideas from BR provide interesting advantages, since they aim to obtain results similar to those humans conclude. An adequate choice of key features and their specification is a first step in order to reconstruct the (complex) phenomena. In Fig. 2, a schema of the main activities aimed at the study of CS is shown, with emphasis on the phenomenological reconstruction phase based on the data available, since this is a fundamental tool for the construction of explanations. The tasks can be grouped into three levels or phases: Reconstruction of the system (modeling), Simulation of the system from the dynamics reconstructed in the first phase, and finally, experimental Validation of the simulated behavior with the real behavior of the system. After the last phase, it is possible to reconsider the reconstruction initially obtained in the first phase to bring another more similar with reality, more accurate model. The theoretical reconstruction of CS could cover only those relevant aspects that are related to the explanation of the phenomenon of interest. It is, therefore, a valuable tool for event explanation.

A phenomenological-based methodology suggested by the application of ideas from BR has been applied to study and simulate CS in previous papers (see, e.g. [14,24]) roughly subsumed in Fig. 1. The first step of the methodology is the selection of relevant attributes from all the available attributes to obtain good predictions, classifications, or explanations on CS. The second step is the use of a sound reasoning method on the selected elements.

Perspectivism as explaining strategy

The application of BR techniques that inherently limit the search for solutions (explanations), outlines a framework where both the search space and, ultimately, the way of seeing the system or event to be explained is implicitly limited by the elaboration of aggregate information from the data. We could say then that a particular perspective is created. This point of view is not free of problems.

If AI engineer assists stakeholders with AI-based systems, what is the plausible explaining acceptable by them? Explanations that focus on a (necessarily) partial view do not necessarily provide the best answer, or even a right answer. Therefore, it is reasonable to think that systems do not offer general solutions to the problem of explanation (or justification of the proposed decision). It is necessary to measure the question from the perspective we are led to by the selective access to (massive) data as well as the inevitable biased selection of dimensions, features and datasets. In our case, the perspective might be strongly based on data for its curation and exploration. This basis compromises the desired independence of the observer's information.

Introduced by Leibniz (and further developed by Nietzsche) Perspectivism starts from the premise that all perception and ideation takes place from a particular perspective (a particular cognitive point of view). Therefore the existence of possible conceptual schemes (perspectives) influencing how the phenomenon is understood and the judgment of its veracity, is assumed. It is important to note that, although it

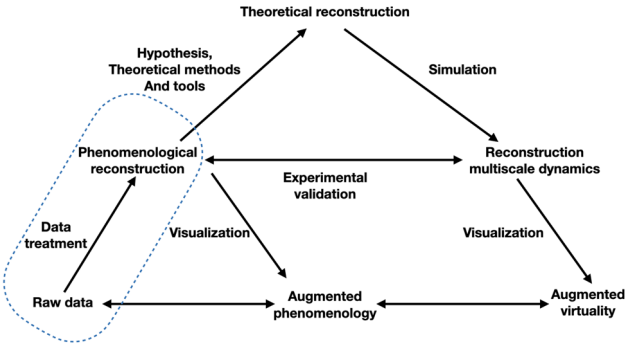

Fig. 2 Tasks in Formal and Applied Epistemology involved in the study of complex systems (from [23])

is assumed that there is no single true perspective to explain the world, it is not necessarily proposed that all perspectives are equally valid.

In a perspectivist view, Science is primarily observer-dependent. Moreover, we see a growing acknowledgment in science studies that all scientific knowledge is perspectival [25]; i.e., that the context established by a scientific discipline is decisive for the kind of observations that can be made. The same phenomenon occurs intra-theory, that is, between different contextual observations sharing the same theory. Hence, it can be concluded that explanations will be in many cases inherently perspectivist artifacts. The perspectivist point of view proposes the existence of many different scientific perspectives with which to analyze a complex problem, all of which can bring value to the study, similar to the fact that a single scientific discipline cannot provide adequate solutions to complex problems. The perspectivist view represents a powerful tool supporting mutual respect and relationship between even very different scientific perspectives [25].

Perspectivism and data curation

The taking of perspective and the use of BR techniques entails in most cases the intended selection of the data for the task of explanation, data curation. In document [26] DARPA agency motivates the focus on Data Science in Explanability Challenge because decisions assisted by BD analytics need from such a selection of which resources will be the study objective to support evidence in their analysis. Such selection could lead to failures or errors that must be analyzed to refine both the procedure and the curation of the content. It is clear that the provision of effective explanations would greatly help with all these tasks [26]. Actually, what might hide Data Curation (particularly selection and interpretation) is the practice of data hermeneutics [27,28]; the entire process is accompanied by information that could become explanations of both, the result and the extraction and curation policies. There is a clear need for this to be documented.

Another risk that comes from data curation is that data bias may lead to the inability to replicate studies, compromising their ability to be reused or generalized as well as the acceptability of the explanation itself [29]. Still, the level of abstraction of the explaining can determine its generalizability, because to consider a too abstract level can compromise both its real understanding and its practical value. Abstractions can simplify explanations, but automating the discovery of abstractions is very challenging (also both sharing your understanding and sharing them) [30]. Such difficulties could lead to a greater gap between scientific rigor and practical relevance. The generalization of explanations, understood as their reusability for several case studies would be strengthened by the availability of more data from multiple sources. This generalization can influence the issue of the preservation of data curation criteria (which could accompany the explanation, since they provide insight). This would also allow the development of richer models and greater understanding. Each model is a reduction of reality and the modeler needs to make choices in light of limited resources for data collection and modeling. However, when more data are available, models often become more complex and too detailed to interpret, unless they possess certain semantic features. For example, instead of trying to confirm the theory through purely deductive approaches, resources such as Linked Data can facilitate the search and analysis of counterfactuals [29], instead of just gathering a representative data sample to confirm our theory.

Perspectivism versus veracity

The adoption of perspectives could affect the veracity (of the explanation given or the model itself). It is adopted here the notion of Veracity as how precise or true a dataset can be. It is referred to as the fidelity of the data concerning the reality that they represent. In the context of Data Science, it takes additional meaning, namely how reliable the data source is, and the confidence in the type and processing of the data. Such aspects need to be studied since they are essential for issues such as avoiding biases, abnormalities, inconsistencies, and others associated with processing such as duplication and volatility. It is a critical issue to be studied in new systems [31], and mandatory if one wants to abandon the idea that ML is data alchemy.

It is particularly interesting the distinction between Veracity in general on one side and the concepts associated with the correctness and validity of the results on the other side. Since databases can be understood at a certain level as models of the definition schemes that govern them, and these are in turn formal theories that represent the universe from which the data are extracted, veracity is very much related to classic problems in Knowledge Engineering (or, if we demand a

certain discipline in the database definition schemes, with the Semantic Web [32]). Therefore, veracity also depends on the quality, safety, accuracy, completeness of the information, etc.

Towards a theoretical general model for explaining phenomena: background

The issue to be addressed now is whether the elements and ideas formerly discussed can be formalized (from a logical, foundational point of view in nature). It should be noted that the aim is to provide a proof of concept, a common framework on which—at least theoretically—one could compare different approaches for explaining. The proposed model is data driven and intended for information fusion from input modules, to meet the conditions outlined in the previous sections. It will be a sort of universal KRR-based (phenomenological) surrogate model. The idea is that any other surrogate model coming from data could be considered immersed within this. Bridging the considerable gap, it is a modest proposal towards a model for XAI similar to what was done in other areas such as ZFC (actually, its so-called Inner Models) that allowed to establish basic elements of Set Theory by means providing a common formal framework.

Our aim is to address the question whether it is possible to consider a universal surrogate model that encompasses any surrogate model and enable the production of acceptable explanations, including those under BR (which limits both options/choices and inferences). It is important to note that both the question (a meta-epistemological question) and the answer (a mathematical approach) are epistemological in nature and based on phenomenological reconstruction philosophy sketched in "Formal epistemology, complex systems and BR". It is not intended here to demonstrate the benefits in practice of this approach, beyond its role as a facilitator of formalizations of some of the issues we have discussed so far (although some of its most important mathematical properties will be demonstrated).

The idea of a universal model for the phenomenological reconstruction of CS is not completely original from the paper (in principle, it can be considered a case of the general strategy of addressing the CS study). The first time it was sketched was in a case study in the specific field of sports betting [14] that clearly illustrated some of the characteristics that the model should have.

The method offers a description of the relationships between the observed data (the basic attributes that can be considered as raw data) and system of logic implications (which can be seen as Horn clauses, which will make it easier to formalize the explanation). The model composed by the elements obtained (a network of concepts and a system of rules) can be considered a surrogate model for explaining qualitative grounded relationships of the System. Mathematical results justify the soundness and completeness of the model concerning the raw data coming from the CS, always from the foundational vision of the problem. The shift to the use of more complex properties should be formalized by what is called formal perspectives (see "The monster context").

In order to present the ideas within a common framework, the notions that have been discussed so far, Formal Concept Analysis (FCA) [19] has been selected. FCA is a mathematical theory for the analysis of qualitative data, hence it is an ideal tool for our purposes Since its origins, with the pioneering works by Rudolph Wille from the 1980s, FCA has experienced an outstanding development in both its theoretical [33] and applied sides [34]. For the reader's convenience, in Table 1 a summary of frequently used notations is presented.

The choice of FCA as basis theory to describe the model is due to, among other reasons, the fact that FCA does not prefix the language of concepts. So, one could start from variables denoting the basic attributes (considering the latter as data coming from perception and output, for example). FCA allows the extraction of concepts, in the following sense: it is mathematized the philosophical understanding of a concept as a unit of thought, comprising its extent and its intent. The extent covers all objects belonging to the concept, and the intent comprises all common attributes valid for all the objects under consideration.

FCA provides algorithms to extract, from data, all units of knowledge with meaning in the sense of the above-mentioned concept notion, as well as it also allows the computation of concept hierarchies from data tables. In short, FCA theory and techniques represents a method for both Data Analysis (organization, exploration, visualization) and Knowledge Retrieval, among other applications. At the computational-logic level, it also provides tools to extract patterns (rules) of behavior from the data and reason with them. One could summarize the reasons that lead us to choose FCA in that two of our objectives resemble two FCA main goals: conceptual clusters extraction (formal concepts endowed with semantic network structure) and data dependencies (implications between attributes) analysis.

Next subsection the basic elements of FCA will be briefly described, as well as the notation that will be used in the rest of the paper, in an effort to make the paper self-contained. The reader is referred to [19] to get both more technical details and a more comprehensive view of FCA.

Formal contexts

The information format used in FCA is organized in an object–attribute table specifying whether an object have an attribute. This table is called Formal Context. It is a three elements set $\mathbb{K} = (G, M, I)$, where G is a non-empty set

Table 1 Frequently used notations throughout the paper

Notation	Definition (and reference)
$\mathbb{K}$	Formal context (Paragraph 41)
$\mathbb{M}$	Monster model (a formal context, Paragraph 56)
$\mathcal{L}$ ($\mathcal{L}_{\mathbb{K}}$)	Implication sets (for $\mathbb{K}$)
$\mathcal{L}_1 \models \mathcal{L}_2$	$\mathcal{L}_2$ is consequence of $\mathcal{L}_1$ (Definition 1)
$\mathcal{L}_1 \equiv \mathcal{L}_2$	$\mathcal{L}_1$ and $\mathcal{L}_2$ are equivalent (Definition 1)
$\lim_i \mathcal{L}_i$	Limit of a sequence $\{\mathcal{L}_i\}_{i\in\mathbb{N}}$ of implication sets (Paragraph 89)
$\mathbb{K}_1 \sim_B \mathbb{K}_2$	$\mathbb{K}_1$ and $\mathbb{K}_2$ share a common B-approximation (Definition 9)
$\mathbb{K}_1 \equiv_B \mathbb{K}_2$	$\mathbb{K}_1$ and $\mathbb{K}_2$ are B-equivalent (Definition 9)
$\mathbb{K}_1 \cong \mathbb{K}_2$	$\mathbb{K}_1$ and $\mathbb{K}_2$ are isomorphic (Definition 10)

Formula 1	River	Coast	Sea
Carp	×		
Escatofagus	×	×	
Bream		×	×
Sparus		×	×
Eel	×	×	×

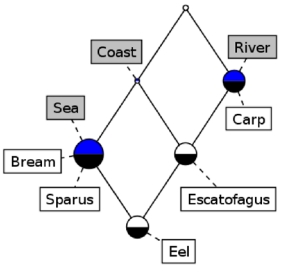

Fig. 3 Formal context on fish, and its associated concept lattice

of objects (events), M is a non-empty set of attributes, and $I \subseteq G \times M$ is a (object–attribute) relation. In the table representation of the formal context, objects and attributes correspond to table rows and columns, respectively, and $(g, m) \in I$ denotes that object g has attribute m. Figure 3 (left) shows a formal context describing fishes (objects) living on different aquatic ecosystems (attributes) is shown. Please note that attributes can be considered as Boolean functions on the set of objects. Any attribute $a \in M$ defines a function $f_a : X \to \mathbb{B}$ where $\mathbb{B} = \{\mathbf{0}, \mathbf{1}\}$ as $f_a(o) = \mathbf{1}$ if and only if $(o, a) \in I$.

A formal context $\mathbb{K} = (G, M, I)$ induces a pair of operators, which we will call here derivation operators. Given $A \subseteq G$ and $B \subseteq M$, they are defined by

$$A' = \{a \in M \mid (o, a) \in I \text{ for all } o \in A\}$$

(that is, the set of attributes shared by all the objects in A) and reciprocally

$$B' = \{o \in G \mid (o, a) \in I \text{ for all } a \in B\}$$

(the objects that have all the attributes of B).

The mathematical instantiation in FCA of the philosophical definition of concept is called formal concept. A formal concept is defined by means of the derivation operators: it is a pair (A, B) of object and attribute sets (called the extent and the intent of the concept, respectively) such that $A' = B$ and $B' = A$. That is, the definition by intention characterizes all the elements that satisfy that definition, and vice versa: the definition by intention contains all the attributes common to those objects. Sometimes concepts are referred to by their intent, which are the so called closed sets. An attribute set B is closed if $B'' = B$ (or equivalently, is the intent of a concept).

Concept lattice

The set of concepts of a context given $\mathbb{K}$ can be endowed with the mathematical structure of lattice, by means of the subconcept relationship. For example, the concept lattice associated with the formal context on fishes of Fig. 3 left is shown in Fig. 3, right[2].

In this representation, each node is a concept, and its intent (extent resp.) is formed by the set of attributes (objects resp.) included along the path to the top (bottom resp.) concept. For example, the bottom concept $(\{eel\}, \{Coast, Sea, River\})$ is a concept that could be interpreted as *euryhaline-fish* (this is not a term of the language represented by attribute set, is something new). This is an example of how FCA does not limit the concepts considered by the chosen language of attributes, and how it induces the discovery of new ones (concepts that do not fit with the extension of any attribute). A more complex example, where the authors analyzed concept lattices about sentiments in social networks [35], in particular on Twitter. The aim was to show that the conceptual structure associated with a large set of aggregated opinions on the same topic can provide an interesting overview of the collective opinion on that subject. From the retrieved conceptual structure it can analyze, at the language level, the evolution of the opinion lexicon in social networks. The work shows how concepts about feelings are not adequately represented with most of the sentiment lexicon used in Social Media [36] arise.

[2] What is represented actually is a Hasse diagram (a graph representing a partial order from bottom to top) induced by the partial order relation among concepts $C_1 \subset C_2$ and there is not any intermediate concept.

An important feature is that basic FCA algorithms extract all concepts from the formal context, which can lead to very complex concept lattices. If a selection of these is desired, more refined algorithms, that focus on the most general or important concepts according to some measure, can be used (see e.g. [37]). The refinement would allow focusing the analysis on an easy-to-handle attribute set, but without losing (as far as possible) the original relations among these. For instance, it is possible to simplify the concept structure but keeping important properties (see, e.g. [38] or [39]). Along with the concept lattice, it is also possible to obtain a KB extracted from the formal context that uses the attributes as representation language using an implication logic.

Implication basis

In FCA, the format of the logical expressions denoting relations among properties (the attributes) is very similar to Horn's clauses. An attribute implication L (over a set M of attributes) is an expression $A \rightarrow B$, where $A, B \subseteq M$. The set of implications on M is denoted by $Imp(M)$.

The semantics of implications is inherited from the natural interpretation of implications in propositional logic but relativized to consider the formal context as the universe of all objects (understanding each set of attributes $\{g\}'$ associated with an object g as an interpretation, that is, the set of true attributes). Formally, it is said that $A \rightarrow B$ is valid for a set T of attributes (or T is a model of the implication), written $T \models A \rightarrow B$, if the following condition is satisfied: *If* $A \subseteq T$ *then* $B \subseteq T$. The implication $A \rightarrow B$ is valid in the context $\mathbb{K} = (G, M, I)$, denoted by $\mathbb{K} \models A \rightarrow B$, if $\{g\}' \models A \rightarrow B$ for any object $g \in G$ (that is to say, the set of attributes of any object in the context formal models the implication). For instance, the implication

$$River, Sea \rightarrow Coast$$

(any fish that lives in both rivers and the sea also live in the coast) is valid within the context of Fig. 3, whilst the implication $River \rightarrow Coast$ does not.

Once semantic truth has been defined, it is possible to formalize the derived notion of entailment.

Definition 1 Let $\mathbb{K} = (G, M, I)$ be a formal context, $\mathcal{L}$ be an implication set and L be an implication. It is said that

1. L follows from $\mathcal{L}$ (or L is consequence of $\mathcal{L}$, denoted by $\mathcal{L} \models L$) if each model (subsets of attributes) of $\mathcal{L}$ also models L. Similarly, it will be written $\mathcal{L} \models \mathcal{L}'$ every implication from $\mathcal{L}'$ is consequence of $\mathcal{L}$.
2. It is said that $\mathcal{L}$ and $\mathcal{L}'$ are equivalent, $\mathcal{L}' \equiv \mathcal{L}$, if $\mathcal{L} \models \mathcal{L}'$ and $\mathcal{L}' \models \mathcal{L}$.

Extending	Need Water	Aquatic	Mobility	Legs
Cat	×		×	×
Leech	×	×	×	
Frog	×	×	×	×
Corn	×			
Fish	×	×	×	

1 < 5 > { } ==> Need water;
2 < 3 > Need water Aquatic ==> Mobility;
3 < 2 > Need water Legs ==> Mobility;

Fig. 4 Context from observation (left) and stem basis (right)

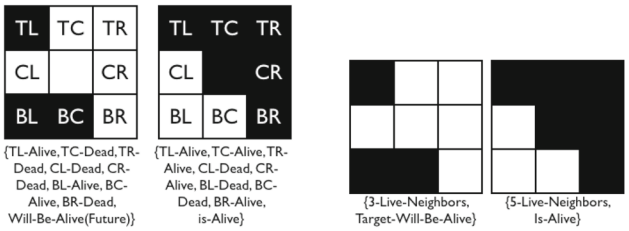

Fig. 5 Two examples of representation of cell state and its neighbour by means the so-called geometric attributes (left) and two examples of representation of cell state and its neighbour by means the attributes induced by Conway's original formulation (right)

3. $\mathcal{L}$ is complete for $\mathbb{K}$ if for every implication L

   If $\mathbb{K} \models L$ then $\mathcal{L} \models L$

4. $\mathcal{L}$ is non-redundant if for each $L \in \mathcal{L}$, $\mathcal{L} \setminus \{L\} \not\models L$.
5. $\mathcal{L}$ is an implication basis for $\mathbb{K}$ if $\mathcal{L}$ is both complete for $\mathbb{K}$ and non-redundant.

The computation of implication bases can be studied from a more general setting, within the field of Lattice theory [40]. A particular basis is the so called Duquenne–Guigues Basis, also called Stem Basis (SB) [41], which is extracted from a type of attribute sets (pseudo-intents) [19]. Figure 4 show a context and its associated Stem basis. Actually, working with complete implication sets would be enough. Regarding implication bases, there exist relatively few solutions to compute them from the formal context. One of the most popular algorithms is Ganter's construction of canonical basis that is a modification of his Next-Closure method for computing concept sets (see e.g. [42] for a discussion on the topic). Stem basis computation, based on pseudo-closed sets computation, also suffers of theoretical complexity barriers. For instance, deciding pseudo-closedness of attribute sets is coNP-complete.

As for any logical implication, forward reasoning is defined in the natural way. The entailment relationship based on the classic production system style will be denoted by $\vdash_p$. In formal terms, without specifying any particular algorithm, the definition that captures the usual rule-firing closure:

Definition 2 Let $\mathbb{K} = (G, M, I)$ and $\mathcal{L} \subseteq Imp(M)$ and $H \subseteq M$. The implicational closure of H, with respect to $\mathcal{L}$, $\mathcal{L}[H]$, is the smallest set $B \subseteq M$ such that:

– $H \subseteq B$
– If there exists $Y_1 \rightarrow Y_2 \in \mathcal{L}$ such that $Y_1 \subseteq B$, then $Y_2 \subseteq B$

Given C a subset of attributes, it will be denoted by $\mathcal{L} \cup H \vdash_p C$ if $C \subseteq \mathcal{L}[H]$.

The logical soundness and completeness, with respect to the entailment, is based on the following result:

Theorem 1 *Let $\mathcal{L}$ be an implication basis for $\mathbb{K}$ and let $\{a_1, \ldots, a_n\} \cup Y$ be a set of attributes of $\mathbb{K}$. The following statements are equivalent:*

1. $\mathcal{L} \cup \{a_1, \ldots a_n\} \vdash_p Y$.
2. $\mathcal{L} \models \{a_1, \ldots a_n\} \rightarrow Y$
3. $\mathbb{K} \models \{a_1, \ldots a_n\} \rightarrow Y$.

The construction of our model for formalizing explaining is based on a series of assumptions that will be introduced when needed or when they can be described. The first one is the following:

Assumption 1 The use of a system of rules (technically, they are definite Horn clauses) enables the construction of explanatory systems.

Note that it is not claimed that every event can be explained, only that we will consider for our model explanations based on information represented by such kind of formulas. Of course, it can be extended.

Semantics for propositional formulas and association rules

Recall that it has already been mentioned that one might consider $\{g\}'$ (being $g \in G$) as an interpretation of propositional logic on language formed by the attributes of M (being $\{g\}'$ the true attributes of such interpretation). Therefore, the validity of any propositional formula can be considered.

Implication basis are sound KBs to be used within a rule-based system in order to reason and also learn. For example, theory and tools from Inductive Logic Programming (ILP) [43] can be applied (as in fact will be done in "Inductive logic programming versus explaining in the model" bellow). Please note, however, that implication basis are designed for entailing only true implications, without any exceptions within the object set nor implications with a low number of counterexamples in the context. Consequently Theorem 1 applies only to valid implications. They should, therefore, also be considered rules with confidence. That is, implications that, while not necessarily logically true, are validated by a significant set of objects. For this purpose, the initial production system must be revised in order to work with confidence [44] as any rule-based system [45], following a relaxed version of Pollock's notion of statistical syllogism [46] (see also [47]). Due to paper length issues, association rules in FCA will not be discussed here.

A toolbox for specifying explanations in complex systems

After introducing the basic elements that will be used in FCA, we return to the reasoning process represented in Fig. 1. It is started with a formal context which contains all the information about the system to study, which comes from all the perceptions/observations that will be objects of a formal context (please note that this would be the closest possible approximation to the event, this digital shadow is the most faithful to the perceptive capacity of the system). The (Boolean) attributes represent everything perceived. Each attribute represents any available data, for example, the ith bit of the temperature representation, the color of the object is red? or the ith bit of the time representation, that is from bit to bit information to indexes that the agent elaborates from information provided by the sensors. Thus a very large set of attributes is available (which could be assumed to be numerable, although if we talk about representing real information that the system receives it is not, it actually is finite). From this set, the values of more descriptive variables (i.e. more complex attributes) can be obtained through computable functions from the available ones. The formal context built from these data will be called Monster Context, and will be denoted by $\mathbb{M}$.[3]

Once $\mathbb{M}$ is considered, the observer (who probably will be the explainer agent) has to select a set of attributes and observations that she/he has considered relevant to study the event (surely, the product of a selection phase following some BR strategy of those described in previous sections). The reasoning is focused then on the formal context induced by that selection (contextual selection) using original attributes, or with new computed attributes (in the latter case the context will be called formal perspective). It is expected reasoning with the induced formal context (represented in Fig. 1 by the box at the bottom right) to explain the event. As discussed, the ideas of the process come from authors' former works on similar strategies [14,49].

The synthesis of simpler formal contexts but with more elaborated (computable) attributes allows the observer to work with aggregated data but of a reasonable size. From the new context (which we will call formal perspective), the observer can focus the study on specific aspects such as the past evolution of that system and/or create hypotheses about

[3] Please note that $\mathbb{M}$ is not an universal context in the sense of [48]. Although it can be considered an immersion within this, $\mathbb{M}$ does not share its fundamental properties.

its future evolution as well as to explain a specific perception with more elaborated concepts and attributes. In [44,49] some technical aspects were detailed. It will be briefly summarized here so that the rest of the article is self-contained and respects a common notation.

Likewise, although the model will be presented using Implication Logic for the formalization of explaining, association rules or more sophisticated probability tools (such as [50–52]) can be used. As stated by Gigerenzer and Goldstein in [53], Probabilistic Mental Models assumes that inferences about unknown states of the world are based on probability cues (see also [54]). It can be said that association rules' confidence extracted from the subcontexts can serve to establish probability cues.

The monster context

Since $\mathbb{M}$ covers all perceivable attributes from events, used or not by the engineer [55], this can be considered as a universal memory from which any other contexts are extracted (corresponding to partial observations, perspectives, or approximations due to perception or information limitations).

Assumption 2 The Monster Context contains all the information on CS available from the observations/perceptions.

Once a specific smaller context is computed from $\mathbb{M}$, it is possible to work with the elements extracted from that, namely concept lattices, implication basis or association rules [55]. Subcontexts of $\mathbb{M}$ can be selected according to BR techniques (the selection of the elements that make up that sub-context and which is in fact a limitation of the solutions search space, see "Some notes on bounded rationality and explaining") to obtain a reasoning system in which it is feasible to predict, analyze or explain events [49,55] (with the obvious limitations from BR). That is to say, concepts of a qualitative nature are drawn from partial data that consider only partial characteristics of the CS, i.e. a partial understanding.

The basic subcontext is one for which it is not necessary to compute new attributes, that is, those of the form

$$\mathbb{K} = (G, M, I_{\mathbb{K}}) \text{ where } G \subseteq \mathbb{O}, M \subseteq \mathbb{A} \text{ and } I_{\mathbb{K}} = \mathbb{I} \cap (G \times M).$$

Given $\mathbb{K}_i = (G_i, M_i, I_i),\ i = 1, 2$ two subcontexts of $\mathbb{M}$, the intersection of $\mathbb{K}_1$ and K_2 is the context

$$\mathbb{K}_1 \cap \mathbb{K}_2 := (G_1 \cap G_2, M_1 \cup M_2, I_1 \cap ((G_1 \cap G_2) \times M_1) \cup I_2 \cap ((G_1 \cap G_2) \times M_2)).$$

Note that this context takes advantage of the values of the attributes of both contexts on the common objects.

In general terms, a way to select a sub-context of $\mathbb{M}$ when we want to study a particular event $o \in \mathbb{O}$ is through what we call contextual selection, formally defined as follows.

Definition 3 Let $\mathbb{M} = (\mathbb{O}, \mathbb{A}, \mathbb{I})$ be the monster context, and let $O \subseteq \mathbb{O}$.

1. A contextual selection on $O \subseteq \mathbb{O}$ and M is a map

$$s : O \to \mathcal{P}(O_1) \times \mathcal{P}(M)$$
$$s(o) = (s_1(o), s_2(o))$$

such that $o \in s_1(o)$

2. A contextual KB for an object o w.r.t. a selection s is an implication basis of $M_{s(o)} := (s_1(o), s_2(o), \mathbb{I} \cap (s_1(o) \times s_2(o)))$

That is, s maps to each o object a sub-context containing o. This way the reasoning will be focused on a subcontext using a selection function on objects and attributes around the event o.

Formal perspectives are contexts built with more elaborated attributes. We will now assume that it has a computability model that outlines the class of computable functions. More precisely, what interests us is the representation of the functions computed by programs as functions on objects (on their attribute values) belonging to sub-contexts. We will not detail this issue (which does not affect the development of the AI part of the model construction).

Definition 4 A computable attribute b on $\mathbb{O}$ is an attribute defined by means of a computable function $f : \mathbb{B}^n \to \mathbb{B}$ and $\{a_1, \ldots a_n\} \subseteq \mathbb{A}$ as

$$b(o) = f(a_1(o), \ldots, a_n(o)).$$

A formal perspective is a context built from M that uses attributes computed from $\mathbb{M}$:

Definition 5 Let $\mathbb{M} = (\mathbb{O}, \mathbb{A}, \mathbb{I})$ a monster context. A formal perspective is a formal context $\mathbb{P} = (G, M, I)$ built from the monster context with a set M of computable attributes.

According to the definition, subcontexts are formal perspectives.

KBs extracted from contextual selections or formal perspectives would be our theoretical model of KL-based surrogate model. Despite its simple data structure, formal contexts are useful structures for Knowledge extraction and reasoning (cf. [19,33,34]).

By considering the interpretation made of the explanation from the monster context as an ILP process (to be considered in "Inductive logic programming versus explaining in the model"), the approach is aligned with the idea of addressing the explaining of (not necessarily emerging) events and concepts from raw data for reasoning in CS.

Argument-based reasoning as a BR-based activity

To make the model explanation more general, available background knowledge B shall be deemed (e.g. in form of propositional logic formulas). This background knowledge would help obtaining or supporting the explanation offered. For example, it can be used to refine the selection of events to those who satisfy B. Also, one could consider B as knowledge shared by both explainer and explainee; information (about the events of the subcontext) known to the explainer or known by explainee (for example, in medical diagnosis [56]). Background knowledge B would be combined with the knowledge extracted from the formal context (implication basis or association rules).

Note that background knowledge B may not be true in $\mathbb{M}$ (for example, due to erroneous or deficient data from sensors, or because $\mathbb{M}$ contains events that are not relevant to the particular problem being studied and therefore do not necessarily have to satisfy B). Also, bear in mind that by its phenomenological nature, this situation is plausible (one does not work with the System but with its digital perception). There exist two options for solving the inconsistency problem. The one chosen here is similar in nature to what would be called existential argumentation (inspired here by Hunter's paper [57]) but by considering sub-contexts rather than subsets of formulas in a knowledge base. In our case that explanation is supported by a contextual selection that models both B and the explanation obtained. Such formal context, the contextual selection/perspective, is what really supports the explanation and thus inconsistency of the implication basis with B is avoided. Therefore, data and sample used for explanation comes from $\mathbb{M}$ (thus answering one of the questions from Paragraph 45). Other option which will not be considered here would be using conservative retraction by means of variable forgetting [58–60].

The idea is that the arguments that will explain the properties of an event will consist of a set of implications plus a subset of the available perceptions, being the set of implications valid in the contextual selection where we work. Therefore, it is interesting to know how the contexts that allow us to extract an explanation behave. The explanation process consists in finding an explanation that implies the attributes of the event under existential argumentation, $\vdash_{\exists}^{B}$ which involves three steps [49]:

1. A question on why an event has a property (attribute a) is raised. On the event (object) some known properties that comes from perceptions o evidences (attribute values) $P = \{a_1, \ldots a_n\}$.
2. A contextual selection outputs a sub-context of $\mathbb{M}$ satisfying—if exists—the available background knowledge B. A contextual KB, $\mathcal{L}$ (in the case of working with association rules, a Luxenburger basis for for some confidence threshold) is computed for the subcontext. Here BR techniques can be very helpful.
3. Learning tools are applied (for example ILP techniques).
4. The result, the explanation, will be the format $H = \langle Y, \mathcal{L}_0 \rangle$ where $Y \subseteq P$, $\mathcal{L} \models \mathcal{L}_0$ and

$$\mathcal{L}_0 \vdash_p Y \rightarrow \{a_0\}.$$

When the process is successful, it will be denoted by

$$\mathbb{M} \vdash_{\exists}^{B} H \rightarrow \{a_0\}.$$

The procedure can be extended to explain other types of information, for example implications (actually $\{a\}$ is in FCA the implication $\emptyset \rightarrow \{a\}$). Taking into account the completeness properties of the bases for its associated context, it will be denoted as follows:

Definition 6 Let L be an implication and B background knowledge.

1. It is said that L is a possible consequence of $\mathbb{M}$ under the background knowledge B, denoted by $\mathbb{M} \models_{\exists}^{B} L$, if there exists $\mathbb{K}$, a nonempty subcontext of $\mathbb{M}$ such that $\mathbb{K} \models B \cup \{L\}$ (so called supporting context).
2. It is said that L is a possible consequence of $\mathbb{M}$ under the background knowledge B for an event $o \in \mathbb{O}$ if the supporting context is induced by a contextual selection for the event o.

Note that, by Theorem 1, when the background knowledge B is a implication set, $\models_{\exists}^{B}$ would be equivalent to $\vdash_{\exists}^{B}$, which is defined by: $\mathbb{M} \vdash_{\exists} L$ if there exists $\mathbb{S} \models B$, a subcontext of $\mathbb{M}$ such that $\mathcal{L}_{\mathbb{S}} \vdash_p L$.

Implication logics do not suffer inconsistency issues. However, the monster context could have incompatible attributes, for example, a pair a_1, a_2 of incompatible attributes verifies that $\neg(a_1 \wedge a_2)$ is true in the environment. When such a formula is included in the background knowledge B it is possible to deal with incompatibility issues, because $\vdash_p$ is an argumentative entailment which works on subcontexts (see classic $\vdash_{\exists}$ in [57]).

To study $\vdash_{\exists}$ under background knowledge, it may be necessary to study the relationship among arguments based on distinct contexts, checking the compatibility of the knowledge implicit in them. A caveat is that compatibility is not assured under background knowledge in any case. For example, let us look at the two compatibility notions associated with the pull back and the push out ones:

Definition 7 Let $M_i = (O_i, A_i, I_i)$, $i = 1, 2$ be two subcontexts of $\mathbb{M}$, and let B be background propositional knowledge on the language of $A_1 \cap A_2$.

– It is said that M_1 and M_2 are upward compatible w.r.t B if there exists a supercontext M of M_1 and M_2 such that $M \models B$.
– It is said that M_1 and M_2 are downward compatible w.r.t B if $M_1 \cap M_2 \models B$.

If two contexts are upward compatible, then they are downward compatible (therefore, event information can be combined through different contextual selections without compromising consistency with background knowledge) but unfortunately, the reciprocal is not true [49] (thus the union of the information of both contextual selections can lead to inconsistencies). It will be seen below how knowledge behaves under continuous extensions of contextual selection.

Inductive logic programming versus explaining in the model

The consideration of explanations as local in nature are a common practice in AI systems, especially those based on deduction. For example, rule-based systems allow, by following the execution trace, extracting an explanation that has two differentiated parts: the rules triggered in the deduction of the particular attribute associated with the event, and the facts of the initial KB that triggered the rules. Therefore, different explanations can be obtained from different executions for the same result. Something similar occurs with recommendation systems, a special case in which the base of facts is the history of previous customers' choices, product valuations, etc. and the rules are those extracted in the data mining process [4].

In the case of FCA, given some observations about a set of attributes, other values can be inferred by executing the production system—implicational closure—associated with an implication basis the format of an explanation for an attribute m will be a pair $H = \langle Y, \mathcal{L}_0 \rangle$ where Y is a set of attributes (that will be observed or assumed by both explainer and explainee, possibly perceptions shared by both) and $\mathcal{L}_0$ is a set of valid implications verifying that $\mathcal{L}_0 \cup Y \models m$ (that is, $\mathcal{L}_0 \vdash_p Y \rightarrow m$). To simplify notation, this fact will be rewritten by $H \models m$. The search for the explanation will be limited to the formal perspective chosen (recall that contextual selections are also formal perspectives). It is, therefore, a sound way to address the complexity of the explanation offered (introduced in Paragraph 5), and would help isolating the beliefs in the hypotheses that conflict with the beliefs of the involved agents. In this way it has also been decided to simplify the notion of explanation so that it is easier to avoid Heuristic Fallacy ¿HACE FALTA? (Paragraph reffallacy) .

A plausible objection to this type of explanation is that even a local explanation may be too complex to be understood without some sort of approximation [4] (in the case of FCA, the complexity of $\mathcal{L}_0$ itself). In this case, the key challenge is to decide what details to leave out in order to create an explanation based on a simple, explanatory model.

Our model shares many characteristics with the version of classic ILP [43] for Propositional Logics. Therefore, core algorithms from ILP can be applied; the general setting for ILP is used here [43,61]. In ILP one starts with some examples (a set of evidences, E), the background theory B, and the hypothesis H. The problem of inductive inference consists, in our case, in ensuring that H behaves as a sufficient knowledge in order to justify the evidence and observing the validity of B.

A set of implications can be rewritten as a logic program, and therefore, Herbrand interpretations can be considered. With that in mind, $\{o\}'$ can be such interpretation for the implication basis. In the case of explaining, it starts with E a subset of the potential explanandum set $\mathcal{E}$, containing all the attributes of $\mathbb{M}$ for which explanations may be requested, plus a set P of attributes that we consider the perceptions from which the explanation process starts, and which are given to the agents (that is, basic perceptions that will not require explanation and thus outside from $\mathcal{E}$). In this case, $E^+ \subseteq \mathcal{E}$ and $E^- \subseteq \mathcal{E} \backslash E^+$, where E^+, E^- are, respectively, the attributes the event has and does not have.

Notions are described and compared with ILP in Table 2. The aim is to find a hypothesis H such that the following conditions, shown in the second column, hold (normal semantics), in the case of ILP, while for the candidate explanation these are shown in the third column. Let us make this idea a little more concrete

Given $H = \langle Y, \mathcal{L}_0 \rangle$, the set $Y + \mathcal{L}_0$ can be considered as a set of defined clauses, there exist an unique minimal Herbrand model $\mathcal{M}(H)$ of $Y + \mathcal{L}_0$. In our case, knowing that every set $\{o\}'$ of the contextual selection actually is a Herbrand model, a classic result of Logic Programming guarantees that there is a unique minimal Herbrand model contained in all, namely the intersection of these. However, in our case, only those induced by objects from the contextual selection $\mathbb{K}$ (each object being one such interpretation) would be used at the intersection. That is, it is the smallest model relativized to $\mathbb{K}$. This model will be denoted by $M_{\mathbb{K}}(H + B)$. In Table 3, ILP under definite semantics and the corresponding FCA-based explaining are compared.

Some aspects of the ILP approach

Please note that the in the description of the general definition framework, any restriction of minimalism or other restrictions on the explanation $H = \langle Y, \mathcal{L}_0 \rangle$ have been excluded in the above definitions. However, this might be desirable in order to produce simpler explanations and, therefore, objective of further extensions. The algorithmic part is not tackled either. However it is interesting to mention that, in our case, variants of classic backward reasoning algorithms for clausal

Table 2 Inductive logic programming (normal semantics) versus FCA-based explaining, using the implicational closure

Notion/condition	ILP	FCA-explaining	Comment
Evidence	$E = E^+ \cup E^-$	$E^+ \subseteq \mathcal{E}$, $E^- \subseteq \mathcal{E} \setminus E^+$, $P \subseteq \mathbb{A} \setminus \mathcal{E}$ (P is the set of perceived attributes)	Event will be in the formal perspective $\mathbb{P}$ (possibly a contextual selection)
Background knowledge	B	B (a set of propositional logic formulas)	Implicational closure $B[.]$ (see "Some notes on bounded rationality and explaining") if B implication set
Hypothesis/explaining	H (Horn clauses set)	$H = \langle Y, \mathcal{L}_0 \rangle$ with $Y \subseteq P$	$\mathbb{K} \models \mathcal{L}_0$, being $\mathbb{K}$ a contextual selection
Prior satisfiability	$B \wedge E^- \not\models \bot$	$B + P \not\models e$ for all $e \in E^-$	$e^- \notin B[P]$ for all $e^- \in E^-$
Posterior satisfiability	$B \wedge H \wedge E^- \not\models \bot$	$B \cup P \cup \mathcal{L}_0 \not\models e^-$ for all $e^- \in E^-$	$e^- \notin [B \cup \mathcal{L}_0](P)$ for all $e^- \in E^-$
Prior necessity	$B \not\models E^+$	$B + P \not\models e^+$ for all $e^+ \in E^+$	$e^+ \notin B[P]$ for all $e^+ \in E^+$
Posterior sufficiency	$B \wedge H \models E^+$	$B \cup Y \cup \mathcal{L}_0 \models \bigwedge E^+$	$e^+ \in (B \cup \mathcal{L}_0)[Y]$ for all $e^+ \in E^+$

Table 3 Inductive logic programming (definite semantics) versus FCA-based explaining

Notion/condition	ILP	FCA-explaining	Comment
Evidence	$E = E^+ \cup E^-$	$E^+ \subseteq \mathcal{E}$, $E^- \subseteq \mathcal{E} \setminus E^+$, $P \subseteq \mathbb{A} \setminus \mathcal{E}$	The work is done in $\mathbb{P} = (G, M, I)$
Background knowledge	B	B (a set of propositional logic formulas)	$\mathbb{P} \models B$
Hypothesis	H (definite clauses)	$H = \langle Y, \mathcal{L}_0 \rangle$ with $Y \subseteq P$	$\mathbb{P} \models \mathcal{L}_0$
Prior satisfiability	$\mathcal{M}(B) \not\models e^-$ for all $e^- \in E^-$	$\mathcal{M}_{\mathbb{P}}(B \cup P) \not\models e^-$ for all $e^- \in E^-$	Please note that $\mathcal{M}(B \cup P) \subseteq \mathcal{M}_{\mathbb{P}}(B \cup P)$
Posterior satisfiability	$\mathcal{M}(B \wedge H) \not\models e^-$ for all $e^- \in E^-$	$\mathcal{M}_{\mathbb{P}}(B \cup Y \cup \mathcal{L}_0) \not\models e^-$ for all $e^- \in E^-$	It can occur that $\mathbb{M} \models e^-$ for some $e^- \in E^-$
Prior necessity	$\mathcal{M}(B + P) \not\models e^+$ for all $e^+ \in E^+$	$\mathcal{M}_{\mathbb{P}}(B + P) \not\models e^+$ for all $e^+ \in E^+$	As above
Posterior sufficiency	$\mathcal{M}(B \cup Y \cup \mathcal{L}_0) \models e^+$ for all $e \in E^+$	$\mathcal{M}_{\mathbb{P}}(B \cup Y \cup \mathcal{L}_0) \models e^+$ for all $e^+ \in E^+$	In the case of FCA can occur that $M \not\models e^+$ for some $e^+ \in E^+$

KB can be applied to an implication basis of $\mathbb{K}$ (based on the evidence we wish to explain) to extract explanations. For example, by modifying diagnostic or other techniques to detect anomalies (cf. [62] Chap. 5 for a general overview) to get the explanation in the required format. Of course, also with ILP techniques

Nevertheless, there exist other approaches to achieve the best explanations, as [63], where authors employ a logical calculus and starting from conditions of a similar nature to those of ILP. Earlier work on this subject was [64], where Josephson and Josephson propose a way of inferring the best explanation as a kind of argument scheme

Simplifying explanations by means of formal perspectives

It should be noted at this point that to simplify the presentation, no formal perspectives have been considered in comparison with ILP. However, the choice of formal perspectives plays a very important role in the intelligibility of the explanation—which is provided by the contextual selection. The reason is that an explanation based on attributes coming from perceptions can be large or cumbersome. Perspectives allow aggregating information in form of attributes understandable by the explainee that can significantly simplify the explanation offered. Let us see an illustrative example, taken from [65], which shows the importance, of the formal perspective selection, in obtaining explanations acceptable to the explainee.

The example is based on the well-known Conway's game of life (GoL). Suppose that in the attribute set M we have the attributes that we will call geometric, that is, those that represent whether each cell in Moore's neighborhood of a given one is alive or dead, $\{Top - Left - Alive, Top - Left - Dead, \ldots, Bottom - Right - Alive, Bottom - Right - Dead\}$ (a total of 18 attributes), plus the attribute that represents whether the cell is alive ($Is - Alive$) or not, $Will - Be - Alive$ (see two examples in Fig. 5). The world from we consider the contextual selection is the monster model corresponding to a grid of 10,000 cells (so 10,000 objects in the associated formal context), which is taken as contextual selection [65]. Using the previous method, it is possible to obtain an explanation of the Conway's game by means of those attributes. This can be done simply by extracting a base of implications and selecting those having in the head of the implication, the attribute relative to the state of the cell studied, which in this case contains more than 700 implications and that are all necessary since they represent essentially different combinations of the environment. Although the shot of rules will always be one of them, if we want to use it to predict the live/dead state in the transition of a cell-object, which is not a readable explanation.

The formal perspective on the same set of objects is now considered, but using the computable attributes from the previous ones, which determine the number of live adjacent cells describing the neighborhood: $\{0 - Live - Neighbors, \ldots, 8 - Live - Neighbors\}$ (note that they are representable by, for example, DNF formulas using the geometric attributes). These are essentially what Conway would use in the original definition, 9 Attributes that are understandable by the explainee, both their definition and the method of their calculation, (see two examples of representation in Fig. 5). The size of the implication base is considerably smaller for the computable attributes (its size is 6). As the attributes are intelligible and the implication set is small, such a base would be considered an acceptable explanation, as opposed to the one built with the raw data from the monster model. In conclusion, note that in that acceptability two issues play a key role: that the formal perspective significantly reduces the number of implications and, most importantly, the attributes used in the perspective are easily computable and intelligible, possessing a simple definition accepted by the explainee (see Fig. 6, where the last implication of the figure is depicted in the concept lattice).

Approximating the monster context and its information

A question to be solved is whether the model allows us to evaluate (theoretically) the security of the explanation, or to study the convergence to a common explanation if we add experience (we extend the contextual selection). Another question is whether there exists any equivalence between contextual selections with full counterfactual information, in the sense that if there is a counterexample for some implication or explanation, the contextual selection contains one. Having the difficulties of compatibility of different contextual selections (Paragraph 74), the study will focus on the continuous extension of a given context. It is being assumed that one works with approximations to knowledge on the CS that could be extracted from $\mathbb{M}$ (either from sub-contexts or perspectives). Therefore, it is necessary to study what happens when more (empirical) pieces of evidence are available, that is to say, when the induced context is increased. The problem will be restricted to the case of subcontexts, and to the following question: If the explanation depends on the contextual selection and this is extended by the experience (i.e. collection of information of events), to what extent we can approach one stable explanation? That is, one would work with formal contexts that are related by the order $\mathbb{K}_1 \subseteq \mathbb{K}_2$ on subcontexts of $\mathbb{M}$, that has to be understood as that $G_{\mathbb{K}_1} \subseteq G_{\mathbb{K}_2}$ and $I_{\mathbb{K}_1} \subseteq I_{\mathbb{K}_2}$ holds.

The formal scenario will consist of a sequence (Fig. 8) of sub-contexts (contextual selections) $\{\mathbb{K}_i\}_{\{i \in I\}}$ where $\langle I, < \rangle$ is a (partial) order and all formal contexts satisfy $B, \mathbb{K}_i \models B$. The knowledge depends on the one hand, on the implication basis of each $\mathbb{K}_i$, $\mathcal{L}_{\mathbb{K}_i}$. On the other hand, it also depends

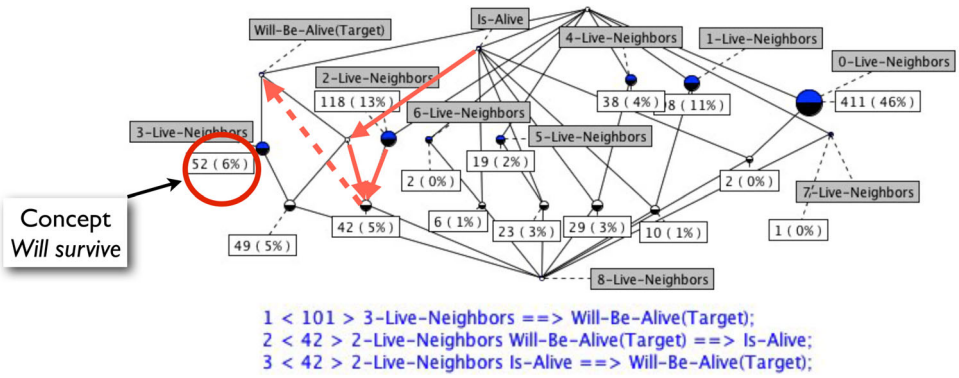

Fig. 6 Concept lattice from Conway's Game of Life for the new perspective. The concept the cell will survive is highlighted and a depiction of the rule 3

Extending	Need Water	Aquatic	Mobility	Legs	Animal	Plant
Cat	×		×	×	×	
Leech	×	×	×		×	
Frog	×	×	×	×	×	
Corn	×					×
Fish	×	×	×		×	
Bacteria	×	×	×		×	×

Fig. 7 Context extending that of Fig. 4

on the behavior of the sequence towards the limit (thinking that this should be a sound approach to the knowledge on the system to study). As mentioned, one is interested in the specific case of incremental observations (each observation adds new items to the subcontext and I the natural numbers), that is, $\mathbb{K}_i \subseteq \mathbb{K}_{i+1}$. The challenge would be to characterize the knowledge from the formal context $\bigcup_i \mathbb{K}_i$, in the expectation of obtaining richer information on the system under observation. If the information on the limit context is not useful, a reconsideration of features will be necessary [15].

There exists a logical characterization of $\mathcal{L}_{\bigcup \mathbb{K}_i}$ that allows focusing the study in desirable features for $\bigcup_i \mathbb{K}_i$. By taking into account that the attribute set can be increased, it is possible to define the limit of bases $\{\mathcal{L}_{\mathbb{K}_i}\}_i$ by means the set of implications defined by

$$L \in \lim \mathcal{L}_{\mathbb{K}_i} \iff \exists i_0 \forall k \geq i_0 \; \mathcal{L}_{\mathbb{K}_k} \models L.$$

The idea is that the value i_0 is related to the point in which there is available information about all attributes from implication L.

Theorem 2 $\mathcal{L}_{\bigcup \mathbb{K}_i} \equiv \lim \mathcal{L}_{\mathbb{K}_i}$.

Proof Let $att(.)$ be the set of attributes that occur in an implication or set of implications, and $\mathbb{K}_i = (G_i, M_i, I_i)$.

$\lim \mathcal{L}_i \models \mathcal{L}_{\bigcup \mathbb{K}_i}$:

Let $L \in \mathcal{L}_{\bigcup \mathbb{K}_i}$. Then $\bigcup \mathbb{K}_i \models L$. Consider i_0 such that $att(L) \subseteq M_i$. Since $\mathbb{K}_{i_0} \subseteq \bigcup \mathbb{K}_i$, it has $\mathbb{K}_{i_0} \models L$ and the same applies to any $\mathbb{K}_j$ with $j \geq i_0$.

$\mathcal{L}_{\bigcup \mathbb{K}_i} \models \lim \mathcal{L}_i$:

Let $L \in lim\mathcal{L}_i$. Let i_0 be such that for any $j \geq i_0$ $\mathbb{K}_j \models L$ (thus $\mathcal{L}_j \models L$).

Let $i_1 \geq i_0$ such that $att(L) \subseteq G_j$. For being a growing succession, for everything $j \geq i_1$ is also true the condition. Using the characterization of Theorem 1, for everything $j \geq i_1$ it has $\mathbb{K}_j \models L$. Since all $o \in \bigcup_i \mathbb{K}_i$ belongs to some $\mathbb{K}_j$ with $j \geq i_1$, then $\{o\}' \models L$. That is, $\bigcup \mathbb{K}_i \models L$, and by the Theorem 1 again $\mathcal{L}_{\bigcup \mathbb{K}_i} \models L$. □

The question now is whether the limit reaches full counterfactual information, in the following sense. It aims that, for any important event of $\mathbb{M}$, that invalidates the conjectured explanation, there is an event in that context with the same properties (concerning the attributes). Therefore, the ideal case of an (incremental) sequence of observation sets should occur when the model $\mathbb{K} = \bigcup \mathbb{K}_i$ satisfies that every relevant type of observation on the system would be represented by an exemplary object. Working with background knowledge and contextual selections has the risk of considering subcontexts that do not necessarily have maximum information. For example, this could happen when the selection chosen to construct the explanation does not contain relevant events to extract explanations consistent with reality.

In our model, we can formalize the idea of formal context with complete relevant information. Theoretically, it is desirable to work with saturated sub-contexts defined as follows. In the next definition, the following notation will be used. Given a set of attributes Y, the propositional formula formed by the conjunction (resp. disjunction) of attributes from Y, will be denoted by $\bigwedge Y$ (resp. $\bigvee Y$). The notion of B-saturation aims to capture the idea that the sub-context contains at least an exemplary event for each possible event that is consistent with B and it is also possible in $\mathbb{M}$. Although the notion would be circumscribed to the language used for explanations, no such restriction will be imposed here in order to avoid complicating the formalization. In addition, it also should be restricted to the events that are of interest for the explanation, and therefore, relative to the sub-language that serves to represent the event. In order to keep the formalization simple, it is supposed to be the whole language, although in each specific problem a much smaller language would be used.

Definition 8 Let $\mathbb{M} = (\mathbb{O}, \mathbb{A}, \mathbb{I})$ be the monster context, B background (propositional) knowledge and $\mathbb{K} = (G, M, I)$ a subcontext. It is said that $\mathbb{K}$ is B- saturated in $\mathbb{M}$ if for every $Y \subseteq M$, if there exists $o \in \mathbb{O}$ such that

$$\{o\}' \models B \cup \bigwedge Y \cup \neg \bigvee (M \setminus Y)$$

then there exists $o_1 \in G$ such that $\{o_1\}'$ also models it.

Please note that it will be assumed, for simplicity, that the attribute language (the context attributes $\mathbb{K}$) will be $\mathbb{A}$.

Proposition 1 *Saturated model exists (although it could be empty)*

Proof Let us consider

$$\mathcal{A} = \{Y \subseteq \mathbb{A} \ : \ \text{exists } o \in \mathbb{O} \ s.t. \ \{o\}' \models B \cup Y \cup \neg \bigvee (M \setminus Y)\}.$$

Then the context $(\bigcup\{Y' \ : \ Y \in \mathcal{A}\}, \mathcal{A}, I)$, where I is the corresponding restriction of $\mathbb{I}$, is B-saturated □

Obtaining a saturated model could be difficult or impossible (in fact, if we consider BR techniques it is almost certain that such a formal context will not be chosen, because of their limitations). Thus a richer contextual selection will be an approximation to the saturated one. It is assumed that, by expanding the selection (e.g. with more events), the context will be closer to the saturated one. The relation between saturated models and approximations to the CS knowledge is studied by means of the so-called B-approximations, to be defined as follows. Another interesting question is what could be done if the $\mathbb{K}$ selected (e.g. using a BR technique) does not satisfy background knowledge. If this were the case, one would be forced to restrict it to a sub-context of $\mathbb{K}$ that validates it. The criterion (BR-based as required) for discarding events to satisfy B should not eliminate those that do. In fact, it is interesting for such sub-context to be maximal with this property. Another question is whether the contextual selection will condition the outcome (explanation). That is to say, it should also be analyzed whether two different conceptual selections can have a common sub-context (or two equivalent sub-contexts) that models B. Let us formalize those notions.

Definition 9 Let $\mathbb{M} = (\mathbb{O}, \mathbb{A}, \mathbb{I})$ be the monster context, B background knowledge and $\mathbb{K}, \mathbb{K}_1, \mathbb{K}_2$ subcontexts of $\mathbb{M}$.

1. A B-approximation of $\mathbb{K}$ is a maximal context of the set

   $$\{\mathbb{S} \subseteq \mathbb{K} \ : \ \mathbb{S} \models B\}.$$

2. $\mathbb{K}_1 \sim_B \mathbb{K}_2$ if both subcontexts share the same B-approximation.

3. $\mathbb{K}_1 \equiv_B \mathbb{K}_2$ (are B equivalent) if there exists $\mathbb{S}_i$ be a B-approximation of $\mathbb{K}_i$ $(i = 1, 2)$ such that their implication bases are equivalent, $\mathcal{L}_{\mathbb{S}_1} \equiv \mathcal{L}_{\mathbb{S}_2}$.

For example, the context from Fig. 7 is a B approximation of that of Fig. 3 for $B = \{(Plant \vee Animal) \wedge \neg(Plant \wedge Animal))\}$

Both B-saturated and B-approximation contexts represent sound approximations to $\mathbb{M}$ that would be obtained through accumulation of events (results of experiments). It is assumed that if relevant experiences are continuously added, the growing sequence of contexts will tend to B-saturation. Likewise, the evolution of the explanation will, therefore, depend on how the contextual selection is extended, that is, on the associated sequence of contexts. It remains to be seen which is the relationship between the different sequences of observations. It is expected that, when the limits are B-saturated, both limits share the same conceptual knowledge about the CS. To see this, we introduce the notion of isomorphism between formal contexts.

Definition 10 Given two subcontexts $\mathbb{K}_i = (G_i, M_i, I_i)$ $(1 \leq 1 \leq 2)$, it is said that $\mathbb{K}_1$ and $\mathbb{K}_2$ are isomorphic, $\mathbb{K}_1 \cong \mathbb{K}_2$, if there exists $F : G_1 \to G_2$ bijective such $\{F(o)\}' = \{o\}'$ for any $o \in G_1$.

That is, for every exemplary event present in one of the contexts there is an event in the other with the same properties (therefore, they share exemplary events).

The following theorem summarizes the above claims. First, the existence of B-approximation is stated. Second, that the approximation of a context determines the information that can be extracted from the context. The third property states that a B-approximation of a saturated context is also B-saturated (that is, it contains at least one event for each event that is compatible with B and possible according to $\mathbb{K}$) and finally the fourth property shows that equivalence can be replaced by isomorphism, under certain basic assumptions and without losing information. The isomorphism relationship between contexts is defined in the mathematical standard way.

Theorem 3 *Assuming $\mathbb{M}$ is countable (or finite) and $\mathbb{K}$ is a subcontext, the following results holds:*

1. *There exists a B-approximation (which can be the empty context) of $\mathbb{K}$*
2. *If $\mathbb{K}_1 \sim_B \mathbb{K}_2$ then $\mathbb{K}_1 \equiv_B \mathbb{K}_2$*
3. *If $\mathbb{K}$ is B-saturated and $\mathbb{S}$ is a B-approximation of $\mathbb{K}$, then $\mathbb{S}$ is also B-saturated.*
4. *Suppose that $\mathbb{K}_1$ and $\mathbb{K}_2$ are B-saturated with $\mathbb{K}_1 \equiv_B \mathbb{K}_2$, and $\mathbb{S}_i$ is B-approximation of $\mathbb{K}_i$ $(i = 1, 2)$. There exist $\mathbb{K}_i^s \subseteq \mathbb{S}_i$ such that*

   $$\mathbb{K}_1^s \cong \mathbb{K}_2^s \text{ and } \mathcal{L}_{\mathbb{S}_i} \equiv \mathcal{L}_{\mathbb{K}_i} \quad (i = 1, 2).$$

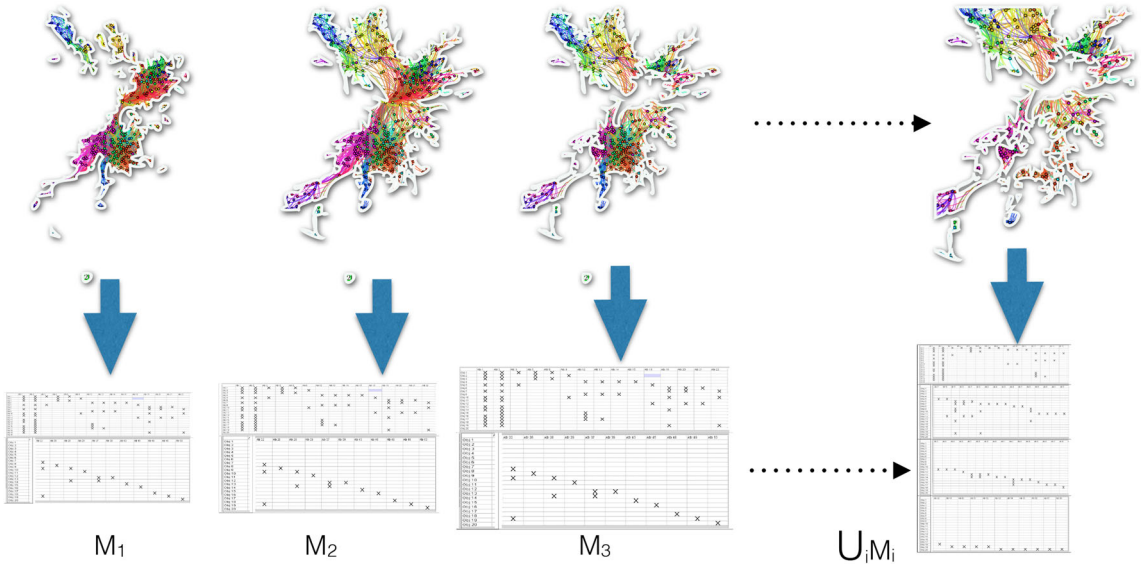

Fig. 8 FCA-based representation of AI-based system evolution

Proof (see Appendix). □

The above result shows how a saturated model approximation (with respect to B) provides relevant and consistent information about the CS since any pair of such models contains a pair of isomorphic contexts with complete information. Thus if one of them contains an exemplary object for a counterexample, the other contains another one that has the same attributes as the first one (i.e., isomorphic). A sequence of observations induces an increasing succession of contexts that accumulate these events and that would approximate a saturated model. In fact, this is desirable as it provides complete information at the limit. The fourth property of the previous theorem would involve that it does not matter what sequence of observations is used if its limit is B-saturated so that the explanation in the format we have proposed would not be essentially different in terms of its logical properties. It is relatively simple to relativize the previous results to a contextual selection or formal perspective, as well as to a restricted language, specialized in the event to explain (Fig. 8).

Representational issues of the approach

The thesis that claims that the overall representation of the perceived information of a complex AI-based system is present in $\mathbb{M}$, could be considered phenomenological in nature. It is not perspectivist but does serve as a basis for the construction of perspectives instead (in fact it should be called micro-perspectives since it comes from micro-phenomenological reconstructions, something like the one used in [66]). The monster context gathers all the information that sensors collect from the environment. It is in this context where BR techniques can be used to select subcontexts to work with, in particular for obtaining explanations. This idea is not new; it shares many similarities with the mechanical-statistical perspective.

The veracity (in the sense of Paragraph 30) is ensured by the universal nature of the Monster Context, and as much as possible from the point of view of the agent who enjoys such perceptions, although it is plainly constrained to the way we understand the mode of perception. If a phenomenological model based on an ongoing process of anticipation and fulfillment was needed (such as that described in [67] for visual perception), certain aspects would need to be reformulated, because only implications are were used.

A different issue occurs when a contextual selection (or a formal perspective) is made. In that case, its intrinsic veracity in the epistemological sense (work is being done with valid implications in the selection but not necessarily in the monster model or a larger selection) is missing. That is to say, it comes from the perspectivist nature of obtaining explanations from sub-contexts. There is also a loss of veracity due to the phenomenological nature of the model. As discussed in Paragraph 69, our model may suffer from inconsistency concerning some background knowledge. Actually, it is one reason why both, B-approximation and B-saturation notions have been studied in "Approximating the monster context and its information". In this way, we have addressed some of the ideas of Paragraph 13. However, this issue should be studied in more detail, especially, its relationship to the so-called Commonsense realism [68].

The mere fact of considering the veracity of the information extracted from the formal perspective could lead the

reader to think that a purely contextualist model is being presented, but this is not the case. Excepting the incompatibility issues addressed in Paragraph 73 (understood as knowledge that is not true but that, due to the use of implications, is not detected, and that was addressed in Paragraph 73), the model intends to be invariant; since the aim is to keep the standards for knowledge invariant with the context, as described in [69]. More specifically, nonskeptical invariantist (the standards for knowledge are relatively low).

Conclusions

The feasibility as well as the formal conditions under which KRR-based systems for explaining observed events/phenomena, produced by Complex (AI-based) Systems, have been investigated. It has been treated in a general way, emphasizing its philosophical, computational, and particularly AI dimensions in the field of Data Science and CS.

The notion of KL-based surrogate model is introduced as a purely theoretical framework. In the case of the FCA-based model presented, explanations are characterized by exploiting standard notions from ILP. In the style of the inner models used in Set Theory or Monster Model in Model Theory, a first formulation of an universal model is proposed to clarify several notions discussed in the first part of the paper.

The theoretical model seeks the generalization of the automated modeling of CS, although it is supposed to be unattainable. The philosophy of our proposal goes towards the presentation of the elements involved in the modelling to represent an idea of explanation and associated notions. In principle, it does not attempt to be a robust basis for automated modelling based on measurements of the observed system variables (as in other fields, as for example [70]).

The reader may recognize in parts of the paper ideas that come from explainability in Expert Systems. These similarities are not surprising since the KBs extracted from formal contexts can be considered as a sort of primitive expert systems. The variable character of the KB built, according to the selected contextual selection or formal perspective where to work, differentiates our model from a classic one. For instance, when choosing a perspective, the explainer could only state that believe the explanation, even if it is true.

The idea of building the structure of concepts and bases of implications from the contextual selection (selection of a point of view in the form of context), lines up somehow with Clancey's hypothesis stated in [71] (p. 109). Clancey conjectured in that paper that human does not retrieve conceptual structures from memory and interpret them; rather, each time we remember, we are constructing a conceptual structure, focusing this way on the impact of situatedness. Also, the overall process of our phenomenological model is according to the task structure for interpretation task suggested by Steels in the same book, [71] (p. 21).

Above theoretical issues contrast with some successful approaches based on the idea. For instance, there are some AI techniques based on BR which allow to overcome obstacles related to the huge size of concept lattices and implication bases associated with complex systems, as these are often backed by big-sized datasets. BR-based methods allow reducing the number of features to be taken into account for achieving specific tasks (e.g. fast and frugal methods [72,73]) Authors have applied this approach in AI tasks, say the evaluation of the available information quality [15], forecasting [14], and reasoning with collective intelligence [74].

An evident limitation of our model, when dealing with the problem of the complete formalization of a complex AI-based system, comes from the tension between what D. Ihde called microperception [75], whose emphasis is on the sensory dimensions—which would correspond to our model—and macroperception, which emphasizes the cultural/hermeneutic dimensions. This latter is closer to the tradition where some systems were defined (which includes concepts already accepted by the community, particular interpretations of some of these, and the adaptation to the knowledge structures that the explainee manages).

There exist approaches sharing aims with our model in the sense of choosing rules for explaining, although their goals are not of foundational nature. For example, Lakkaraju et al. [76] propose an approximation algorithm to generate global explanations in the form of small compact rule sets, each of which captures a certain behavior of the black box model under certain conditions. Another related work is that of Ribeiro et al. [77] where the authors describe an optimization algorithm that balances accuracy and coverage in the search for rules. In the recent work [78] the authors apply thresholds to obtain logical descriptions—in the form of Boolean functions or probabilistic rules from trained neural networks—using two different methods, which add to other approaches already known (see the introduction of the mentioned paper for references to previous approaches). Their proposal is endogenous in nature; it considers the internal objects (neurons) of the system. If one wanted to interpret this approach in our proposal, the appropriate attributes to interpret it as a formal perspective need to be designed, although the problem selecting a sound threshold would be persistent. In FCA, the use of thresholds to disaggregate non-Boolean attributes (e.g., with scales or multivalued) is a common practice. However, the adequate choice of them depends on the problem (see e.g. [14]).

The epistemological variety of explanations is a psychological factor (and prospective object of study in BR) that can strengthen explainee's confidence in the explained hypothesis [79]. In our model, it is assumed that the explainer might obtain several explanations, which may or may not be essen-

tially different. The variety of explanations can be considered in our model at two levels, namely as a variety of explanations in a contextual selection given and as a variety of contexts that validate an explanation (even a combination of both). It would also be affected by the BR-based selection techniques. We have not discussed this topic here. According to Landes [79] that intuitive idea is not sound in general, it must be handled with care.

Future work

The choice of implication logic as a foundational support for explanations benefits both from FCA's results as well as from the approximate nature of Horn's clauses in the face of counterfactual issues asked by explainee [80], which we will consider in a future work. However, our proposal is only a first proof of concept towards a formalization of the notions involved in AI-assisted explaining of events (or justification of decisions) of CS (always keeping in mind that a number of sophisticated AI-based systems are actually CS).

Roughly speaking, FCA has been applied to two different levels for reasoning on CS. In the micro-dimension, the specification is intuitive, and it does not need to reduce the information size. In the macro dimension, when considering the overall knowledge, perspectivist methods (contextual selections, formal perspectives, approximations to saturated models, etc.) should play an important role. For building explanations, the paper is focused on implication basis, however several results can be extended or adapted to association rules. They will have the natural particularities and this will be aim of future work. Likewise, further research will focus on the way of estimating of approximations fitness.

Selecting and reasoning with contextual selection may be analyzed as an argument-based approximation but for reasoning on the system (under certain semantic constraints to be studied). Something like the conflict-free arguments [57] would provide useful insights about how humans reason and act within CS. This topic will be the subject of future research.

It is important to recall that we do not deal specifically with causality here. We have tried to discuss the strengths of KRR for enhancing XAI in Data Science, without going into the nature of the explanation that would be offered. However, it would not be difficult to adapt Halpern's (and Pearl) framework for causality [81], to decide whether the explanation offered in our model is in accordance with that definition. This question will be addressed in future work, individually or together with the nature of the explanation, according to Lewis et al. [22] classification (see Paragraph 21).

Finally, although the paper concerns on foundational issues and not on pragmatics, works referenced in Paragraph 106 show that approaches similar to that of the paper, can provide practical solutions in XAI. The practical benefits of the proposed framework in real life scenarios (in the line of the cited [14]) will be further explored.

Funding This work is supported by Agencia Estatal de Investigación, project PID2019-109152GB-I00/AEI/10.13039/501100011033.

Data availability No datasets have been used for this work.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Code availability No specific software has been used in this paper.

Appendix

Proof of Theorem 3

Proof items (1) and (2) are trivial.

The proof of (3) is based on the fact that if $o \in O_{\mathbb{K}}$ and $\{o\}' \models B$, then the context induced by adding o to the context $\mathbb{S}$ would also be a model of B. Since $\mathbb{S}$ is maximal, then $o \in O_{\mathbb{S}}$

(4) (sketch): By (3), $\mathbb{K}_1$ and $\mathbb{K}_2$ are both B-saturated contexts. $\mathbb{K}_1^s$ and $\mathbb{K}_2^s$ are built as follows: suppose $\mathbb{O}_{\mathbb{S}_i} = \{o_n^i\}_n$.

A morphism is defined as $F = \bigcup F_k$, where $F_k = \{(c_i, d_i) \ : \ i \leq k\}$ is an increasing sequence, defined by recursion as follows:

- $i = 0$: Let $c_0 = o_0$. Since $o_0 \models \{o_0\}' + B + \neg \bigvee \{A \ : \ A \notin \{o\}'\}$ and $\mathbb{K}_2$ is B saturated, there exists $b \in \mathbb{K}_2$ such that

$$\{b\}' \models \{o_0\}' + B + \neg \bigvee \{A \ : \ A \notin \{o_0\}'\}$$

Since $\mathbb{S}_2$ is a B-approximation of $\mathbb{K}_2$, and $(O_{\mathbb{S}_2} \cup \{b\}, A_{\mathbb{S}_2})$ induces a subcontext of $\mathbb{K}_1$ model of B, then by the maximality of $\mathbb{S}_2$ there exists k in $\mathbb{S}_2$ such that $\{o_k^2\}' = \{b\}' = \{o_0^1\}'$. Then let $d_0 := o_k^2$.

– Suppose defined F_k and define (c_{k+1}, d_{k+1}). Let

$$m = \min\{n : o_n \notin \{c_0, \ldots, c_k\} \text{ and } \{o_n\}' \neq \{c_j\}' \text{ for any } j \leq k\}.$$

If there is not such m, the procedure stops and $F = F_k$. Otherwise the procedure to select $c_{k+1} = o_m^1$ and the selection procedure of d_{k+1} is applied from c_{k+1} as in the previous item.

Once the procedure is finished, the contexts are defined from the domain and the image of F;

$$\mathbb{K}_1^s := (dom(F), \mathbb{A}, \mathbb{I} \cap (dom(F) \times \mathbb{A}))$$
$$\mathbb{K}_2^s := (rang(F), \mathbb{A}, \mathbb{I} \cap (rang(F) \times \mathbb{A})).$$

Thus, it is clear that $\mathbb{K}_1^s \cong \mathbb{K}_2^s$. As it is verified that $\mathcal{L}_{\mathbb{S}_1} \equiv L_{\mathbb{S}_2}$, it would be enough to prove that $\mathcal{L}_{\mathbb{S}_1} \equiv \mathcal{L}_{\mathbb{K}_1^s}$.

Of course $\mathcal{L}_{\mathbb{K}_1^s} \models \mathcal{L}_{\mathbb{S}_1}$, because $\mathbb{K}_1^s \subseteq \mathbb{S}_1$ and, therefore, all implication L of $\mathcal{L}_{\mathbb{S}_1}$ is valid in the context $\mathbb{K}_1^s$. Then for completeness of the base, it is true that $\mathcal{L}_{\mathbb{K}_1^s} \models L$.

To demonstrate the symmetrical relationship, $\mathcal{L}_{\mathbb{S}_1} \models \mathcal{L}_{\mathbb{K}_1^s}$, let us suppose by reductio ad absurdum that $\mathcal{L}_{\mathbb{S}_1} \not\models L$ for some $L \in \mathcal{L}_{\mathbb{K}_1^s}$. Then there exists o_k^1 an object in $\mathbb{S}_1$ such that $\{o_k^1\}' \not\models L$. This object cannot belong to $\mathbb{K}_1^s$ since it is a valid implication in this one. However, the set $\{o_k^1\}$ must have been considered in the construction of F. So there exists o_j^1 such that $\{o_j^1\}' = \{o_k^1\}'$ which belongs to the F domain, and therefore belongs to $\mathbb{K}_1^s$, so then $\mathbb{K}_1^s \not\models L$, a contradiction. □

References

1. Arrieta A Barredo, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115
2. Doran D, Schulz S, Besold TR (2017) What does explainable AI really mean? A new conceptualization of perspectives. In: Besold TR, Kutz O (eds) Proc. first int. workshop on comprehensibility and explanation in AI and ML, vol 2071 of CEUR workshop proceedings, CEUR-WS.org, pp 1–8
3. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38
4. Weld DS, Bansal G (2018) Intelligible artificial intelligence. CoRR arXiv:1803.04263
5. Booth S, Muise C, Shah J (2019) Evaluating the interpretability of the knowledge compilation map: communicating logical statements effectively. In: Kraus S (ed) Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, Macao, China, Aug 10–16, 2019, ijcai.org, pp 5801–5807
6. Li J, Liu H (2017) Challenges of feature selection for big data analytics. IEEE Intell Syst 32(2):9–15
7. Weld DS, Bansal G (2019) The challenge of crafting intelligible intelligence. Commun ACM 62(6):70–79
8. Bornstein, A. Is artificial intelligence permanently inscrutable? Nautilus. September 1, 2016; http://nautil.us/issue/40/learning/is-artificialintelligence-permanently-inscrutable
9. Townsend J, Chaton T, Monteiro JM (2019) Extracting relational explanations from deep neural networks: a survey from a neural-symbolic perspective. IEEE Trans Neural Netw Learn Syst 31(9):3456–3470
10. Newell A (1982) The knowledge level. Artif Intell 18(1):87–127
11. Davis R, Shrobe H, Szolovits P (1993) What is a knowledge representation? AI Mag 14(1):17
12. Addis T (2014) Natural and artificial reasoning - an exploration of modelling human thinking, advanced information and knowledge processing. Springer, Berlin
13. Forrester AIJ, Sobester A, Keane AJ (2008) Engineering design via surrogate modelling - a practical guide. Wiley, Hoboken
14. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2013) Complex concept lattices for simulating human prediction in sport. J Syst Sci Complex 26(1):117–136
15. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2013) On the phenomenological reconstruction of complex systems–the scale-free conceptualization hypothesis. Syst Res Behav Sci 30(6):716–734
16. Aranda-Corral GA, 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. Springer International Publishing, Cham, pp 243–260
17. Aranda-Corral GA, Díaz JB, Páez JG (2015) Towards a soft evaluation and refinement of tagging in digital humanities. In: 10th International conference on soft computing models in industrial and environmental applications. Springer International Publishing, Cham, pp 79–89
18. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2014) Simulating language dynamics by means of concept reasoning. In: Di Caro GA, Theraulaz G (eds) Bio-inspired models of network, information, and computing systems. Springer International Publishing, Cham, pp 296–311
19. Ganter B, Wille R (1997) Formal concept analysis: mathematical foundations, 1st edn. Springer-Verlag, New York
20. Weiskopf DA (2015) Observational concepts. The MIT Press, Cambridge, pp 223–248 (**Ch. 9**)
21. Simon HA (1957) Models of man: social and rational : mathematical essays on rational human behavior in a social setting, continuity in administrative science. Ancestral Books in the Management of Organizations, Garland Publishing, Incorporated
22. Lewis RL, Howes AD, Singh S (2014) Computational rationality: linking mechanism and behavior through bounded utility maximization. Top Cogn Sci 6(2):279–311
23. Bourgine P, Chavalarias D, Perrier E (2009) The CSS roadmap for the science of complex systems. Tech. rep., ASSYST, Paris
24. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2013) Qualitative reasoning on complex systems from observations. Hybrid artificial intelligent systems. Springer, Berlin, pp 202–211
25. Alrøe HF, Noe E (2014) Second-order science of interdisciplinary research: a polyocular framework for wicked problems. Constr Found 10(1):65–76
26. DARPA (2016) Explainable artificial intelligence (XAI) program. Tech. rep., Defense Advanced Research Projects Agency. http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf
27. Romele A, Severo M, Furia P (2020) Digital hermeneutics: from interpreting with machines to interpretational machines. AI Soc 35:73–86
28. Gerbaudo P (2020) From data analytics to data hermeneutics. Online political discussions, digital methods and the continuing relevance of interpretative approaches. Digit Cult Soc 2(2):95–112

29. Janssen M, Kuk G (2016) Big and open linked data (bold) in research, policy, and practice. J Organ Comput Electron Commer 26(1–2):3–13
30. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z (2019) XAI–explainable artificial intelligence. Sci Robot 4(37). https://doi.org/10.1126/scirobotics.aay7120
31. Aa V (2015) The field guide to data science, 2nd edn. Booz Allen Hamilton, McLean
32. Alonso-Jiménez JA, Borrego-Daz J, Chávez-González AM, Martín-Mateos FJ (2006) Foundational challenges in automated semantic web data and ontology cleaning. IEEE Intell Syst 21(1):42–52
33. Poelmans J, Kuznetsov SO, Ignatov DI, Dedene G (2013) Review: formal concept analysis in knowledge processing: a survey on models and techniques. Expert Syst Appl 40(16):6601–6623
34. Poelmans J, Ignatov DI, Kuznetsov SO, Dedene G (2013) Formal concept analysis in knowledge processing: a survey on applications. Expert Syst Appl 40(16):6538–6560
35. Borrego-Díaz J, Galán-Páez J (2014) Discovering new sentiments from the social web. arXiv preprint arXiv:1407.0374
36. Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 28(2):15–21
37. Stumme G, Taouil R, Bastide Y, Pasquier N, Lakhal L (2002) Computing iceberg concept lattices with titanic. Data Knowl Eng 42(2):189–222
38. Shao M-W, Guo Y-L (2008) Attribute reduction of large crisp-real concept lattices. In: 2008 International conference on machine learning and cybernetics, vol 1, pp 395–400
39. Dias SM, Vieira NJ (2015) Concept lattices reduction. Expert Syst Appl 42(20):7084–7097
40. Caspard N, Monjardet B (2003) The lattices of closure systems, closure operators, and implicational systems on a finite set: a survey. In: Discrete Applied Mathematics, vol 127, issue 2, pp 241–269, ordinal and symbolic data analysis (OSDA '98), Univ. of Massachusetts, Amherst, Sep 28–30, 1998
41. Guigues JL, Duquenne V (1986) Familles minimales d'implications informatives résultant d'un tableau de données binaires. Math Sci Hum 95:5–18
42. Obiedkov SA, Duquenne V (2007) Attribute-incremental construction of the canonical implication basis. Ann Math Artif Intell 49(1–4):77–99
43. Muggleton S, de Raedt L (1994) Inductive logic programming: theory and methods. J Log Program 19:629–679 (**Special Issue: Ten Years of Logic Programming**)
44. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2011) Confidence-based reasoning with local temporal formal contexts. In: Proceedings of the 11th international conference on artificial neural networks conference on advances in computational intelligence - volume part II, IWANN'11, Springer-Verlag, Berlin, pp 461–468
45. Giarratano JC, Riley GD (2005) Expert systems: principles and programming. Brooks/Cole Publishing Co., Pacific Grove
46. Pollock J (1995) Cognitive carpentry: a blueprint for how to build a person, Bradford Bks. MIT Press, Cambridge
47. Pollock J (2006) Thinking about acting: logical foundations for rational decision making. Oxford University Press, Oxford
48. Levental SM (2012) Study of a universal formal context. Sib Math J 53(5):810–820
49. Aranda-Corral GA, Borrego-Díaz J, Galán-Péz J (2011) Bounded rationality for data reasoning based on formal concept analysis. In: Proceedings of the 2011 22nd international workshop on database and expert systems applications, DEXA'11, IEEE Computer Society, Washington, DC, USA, pp 350–354
50. Oberstone J (2009) Differentiating the top English premier league football clubs from the rest of the pack: identifying the keys to success. J Quant Anal Sports 5(3):1–29
51. Min B, Kim J, Choe C, Eom H, McKay RIB (2008) A compound framework for sports results prediction: a football case study. Knowl Based Syst 21(7):551–562
52. Carmichael F, Thomas D, Ward R (2000) Team performance: the case of English premiership football. Manag Decis Econ 21(1):31–45
53. Gigerenzer G, Goldstein DG (1996) Reasoning the fast and frugal way: models of bounded rationality. Psychol Rev 103(4):650–669
54. Brunswik E (1955) Representative design and probabilistic theory in a functional psychology. Psychol Rev 62(3):193–217
55. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2011) Selecting attributes for sport forecasting using formal concept analysis. CoRR arXiv:1107.5474
56. Stanley DE, Campos DG (2013) The logic of medical diagnosis. Perspect Biol Med 56(2):300–315
57. Hunter A (2000) Reasoning with inconsistency in structured text. Knowl Eng Rev 15(4):317–337
58. Alonso-Jiménez JA, Aranda-Corral GA, Borrego-Díaz J, Fernández-Lebrón MM, Hidalgo-Doblado M (2018) A logic-algebraic tool for reasoning with knowledge-based systems. J Log Algebr Methods Program 101:88–109
59. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2020) A model of three-way decisions for knowledge harnessing. Int J Approx Reason 120:184–202
60. Aranda-Corral G.A., Borrego-Díaz J., Galán-Páez J., Caballero A.T. (2019) On Experimental Efficiency for Retraction Operator to Stem Basis. In: Cornejo M., Kóczy L., Medina J., De Barros Ruano A. (eds) Trends in Mathematics and Computational Intelligence. Studies in Computational Intelligence, vol 796. Springer, Cham. https://doi.org/10.1007/978-3-030-00485-9_8
61. Muggleton S (1995) Inverse entailment and Progol. New Generat Comput 13(3 & 4):245–286
62. Poole D, Mackworth AK (2010) Artificial Intelligence - foundations of computational agents. Cambridge University Press, Cambridge
63. Millson J, Straßer C (2019) A logic for best explanations. J Appl Nonclassical Log 29(2):184–231
64. Josephson J, Josephson S (1996) Abductive Inference: Computation, Philosophy, Technology, Cambridge University Press
65. Aranda-Corral GA, Borrego-Díaz J, Galán-Páez J (2013) Qualitative reasoning on complex systems from observations. In: Pan J-S, Polycarpou MM, Woźniak M, de Carvalho ACPLF, Quintián H, Corchado E (eds) Hybrid Artif Intell Syst. Springer, Berlin, pp 202–211
66. Moguillansky C, Vásquez-Rosati A (2019) An analysis procedure for the micro-phenomenological interview. Constr Found 14:123–145
67. Madary M (2017) Visual phenomenology. The MIT Press, Cambridge
68. Pappas GS (2018) Berkeley's thought. Cornell University Press, Ithaca
69. Hemp D (2010) 1 knowledge and conclusive evidence. In: Camp-1880 bell JK, 'Rourke MO, Silverstein HS (eds) Knowledge and 1881 skepticism. MIT Press, Cambridge, p 27–43
70. Tanevski J, Todorovski L, Džeroski S (2020) Combinatorial search for selecting the structure of models of dynamical systems with equation discovery. Eng Appl Artif Intell 89:103423
71. Steels L, McDermott J (eds) (1994) The knowledge level in expert systems. Conversations and commentary, vol 10 of perspectives in artificial intelligence, Academic Press
72. Goldstein DG, Gigerenzer G (2009) Fast and frugal forecasting. Int J Forecast 25(4):760–772

73. Andersson P, Ekman M, Edman J (2003) Forecasting the fast and frugal way: a study of performance and information-processing strategies of experts and non-experts when predicting the world cup 2002 in soccer. Working Paper Series in Business Administration 2003:9, Stockholm School of Economics
74. Galán-Páez J, Borrego-Díaz J, de Miguel-Rodríguez J (2015) Extracting emergent knowledge about the socioeconomic urban contexts. In: Adjunct proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2015 ACM international symposium on wearable computers, UbiComp/ISWC'15 Adjunct, Association for Computing Machinery, New York, NY, USA, pp 1571–1574
75. Ihde D (1995) Postphenomenology: essays in the postmodern context. In: Northwestern University studies in phenomenology & existential philosophy. Northwestern University Press
76. Lakkaraju H, Kamar E, Caruana R, Leskovec J (2017) Interpretable & explorable approximations of black box models. CoRR arXiv:1707.01154
77. Ribeiro MT, Singh S, Guestrin C (2018) Anchors: high-precision model-agnostic explanations. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18). AAAI Press, pp 1527–1535
78. Liu P, Melkman AA, Akutsu T (2020) Extracting boolean and probabilistic rules from trained neural networks. Neural Netw 126:300–311
79. Landes J (2020) Variety of evidence and the elimination of hypotheses. Eur J Philos Sci 10(2):12
80. Arias M, Balcázar JL, Tirnauca C (2017) Learning definite horn formulas from closure queries. Theor Comput Sci 658(PB):346–356
81. Halpern JY (2016) Actual causality. The MIT Press, Cambridge