Philosophy & Technology (2024) 37:88
https://doi.org/10.1007/s13347-024-00775-3

RESEARCH ARTICLE

On the Nature of Explanation: An Epistemological-Linguistic Perspective for Explanation-Based Natural Language Inference

Marco Valentino[1] · André Freitas[1,2]

Received: 9 January 2024 / Accepted: 21 June 2024

Abstract
One of the fundamental research goals for explanation-based Natural Language Inference (NLI) is to build models that can reason in complex domains through the generation of *natural language explanations*. However, the methodologies to design and evaluate explanation-based inference models are still poorly informed by theoretical accounts on the nature of explanation. As an attempt to provide an epistemologically grounded characterisation for NLI, this paper focuses on the scientific domain, aiming to bridge the gap between theory and practice on the notion of a *scientific explanation*. Specifically, the paper combines a detailed survey of the modern accounts of scientific explanation in Philosophy of Science with a systematic analysis of corpora of natural language explanations, clarifying the nature and function of explanatory arguments from both a top-down (categorical) and a bottom-up (corpus-based) perspective. Through a mixture of quantitative and qualitative methodologies, the presented study allows deriving the following main conclusions: (1) Explanations cannot be entirely characterised in terms of *inductive* or *deductive* arguments as their main function is to perform *unification*; (2) An explanation typically cites *causes* and *mechanisms* that are responsible for the occurrence of the event to be explained; (3) While natural language explanations possess an intrinsic causal-mechanistic nature, they are not limited to causes and mechanisms, also accounting for pragmatic elements such as *definitions*, *properties* and *taxonomic relations*; (4) Patterns of *unification* naturally emerge in corpora of explanations even if not intentionally modelled; (5) Unification is realised through a process of *abstraction*, whose function is to provide the inference mechanism for subsuming the event to be explained under recurring patterns and high-level regularities. The paper contributes to addressing a fundamental gap in classical theoretical accounts on the nature of scientific explanations and their materialisation as linguistic artefacts. This characterisation can support a more principled design and evaluation of explanation-based AI systems which can better interpret, process, and generate natural language explanations.

Extended author information available on the last page of the article

Published online: 15 July 2024

Keywords Scientific explanation · Natural language explanations · Natural language inference

1 Introduction

Building models capable of performing complex inference through the generation of *natural language explanations* represents a fundamental research goal for explainability in Artificial Intelligence (AI) (Došilović et al., 2018; Danilevsky et al., 2020; Thayaparan et al., 2020). However, while current lines of research focus on the development of explanation-based models and benchmarks (Dalvi et al., 2021; Jansen et al., 2018; Jhamtani & Clark, 2020; Thayaparan et al., 2021b; Wiegreffe & Marasovic, 2021; Xie et al., 2020), the applied methodologies are still poorly informed by formal accounts and discussions on the nature of explanation (Cabrera, 2021; Erasmus & Brunet, 2022; Miller, 2018b; Prasetya, 2022; Tan, 2021; Woodward et al., 2017). When describing natural language explanations, in fact, existing work rarely recur to formal characterisations of what constitutes an *explanatory argument*, and are often limited to the indication of generic properties in terms of *supporting evidence* or *entailment* relationships (Camburu et al., 2018; Dalvi et al., 2021; Valentino et al., 2021a; Yang et al., 2018). Bridging the gap between theory and practice, therefore, can accelerate progress in the field, providing new opportunities to formulate clearer research objectives and improve the existing evaluation methodologies (Camburu et al., 2020; Clinciu et al., 2021; Jansen et al., 2021; Valentino et al., 2021a).

As an attempt to provide an epistemologically grounded characterisation for explanation-based Natural Language Inference (NLI), this paper investigates the notion of *scientific explanation* (Salmon, 1984, 2006), studying it as both a *formal object* and as a *linguistic expression*.

To this end, the paper is divided in two main sections. The first part represents a systematic survey of the modern discussion in Philosophy of Science around the notion of a scientific explanation, shedding light on the nature and function of explanatory arguments and their constituting elements (Hempel & Oppenheim, 1948; Kitcher, 1989). Following the survey, the second part of the paper presents a corpus analysis aimed at qualifying sentence-level *explanatory patterns* in corpora of natural language explanations, focusing on datasets used to build and evaluate explanation-based inference models in the scientific domain (Jansen et al., 2014; Xie et al., 2020).

Overall, the paper presents the following main conclusions:

1. **Explanations cannot be exclusively characterised in terms of *inductive* or *deductive* arguments.** Specifically, the main function of an explanation is not of *predicting* or *deducing* the event to be explained (*explanandum*) (Hempel, 1965), but the one of showing how the explanandum fits into a *broader underlying regularity*. This process is known as *unification*, and it is responsible for the creation

of *explanatory patterns* that can account for a large set of phenomena (Friedman, 1974; Kitcher, 1981).
2. **An explanation typically cites part of the causal history of the explanandum**, fitting the event to be explained into a *causal nexus* (Salmon, 1984). There are two possible ways of constructing causal explanations: (1) an explanation can be *etiological* – i.e., the explanandum is explained by revealing part of its causes – or (2) *constitutive* – i.e., the explanation describes the underlying mechanism giving rise to the explanandum. Evidence of this feature is empirically found in the corpus analysis, which reveals that the majority of natural language explanations, indeed, contain references to mechanisms and/or direct causal interactions between entities (Jansen et al., 2014).
3. **While explanations possess an intrinsic causal-mechanistic nature, they are not limited to causes and mechanisms.** In particular, additional knowledge categories such as *definitions*, *prop- erties* and *taxonomic relations* seem to play an equally important role in building an explanatory argument. This can be attributed to both *pragmatic aspects* of natural language explanations as well as inference mechanisms supporting abstraction and *unification*.
4. **Patterns of unification naturally emerge in corpora of explanations.** Even if not intentionally modelled, *unification* seems to be an emergent property of corpora of natural language explanations (Xie et al., 2020). The corpus analysis, in fact, reveals that the distribution of certain explanatory sentences is connected to the notion of *unification power* and that it is possible to draw a parallel between inference patterns emerging in natural language explanations and formal accounts of explanatory unification (Kitcher, 1989).
5. **Unification is realised through a process of abstraction.** Specifically, abstraction represents the fundamental inference mechanism supporting unification in natural language, connecting concrete instances in the explanandum to high-level concepts in central explanatory sentences. This process, realised through specific linguistic elements such as definitions and taxonomic relations, is a funda- mental part of natural language explanations, and represents what allows subsuming the event to be explained under high-level patterns and unifying regularities.

We conclude by suggesting how the presented findings can open new lines of research for explanation- based AI systems, informing the way the community should approach model creation and the design of evaluation methodologies for natural language explanations.

The paper contributes to addressing a fundamental gap in classical theoretical accounts on the nature of scientific explanations and their materialisation as linguistic artefacts. This characterisation can support a more principled design of AI systems that can better interpret and generate natural language explanations. To the best of our knowledge, while previous work on natural language explanations have performed

quantitative and qualitative studies in terms of knowledge reuse and inference categories (Jansen, 2017; Jansen et al., 2016), this study is the first to explore the relation between linguistic aspects of explanations and formal accounts in

Philosophy of Science (Woodward et al., 2017), providing a unified epistemological-linguistic perspective for the field.

2 Scientific Explanation: The Epistemological Perspective

The ultimate goal of science goes far beyond the pure prediction of the natural world. Science is constantly seeking a deeper understanding of observable phenomena and recurring patterns in nature by means of scientific theories and explanations. Most philosophers define an explanation as an answer to a *"why"* question, aiming at identifying and describing the reason behind the occurrence and manifestation of particular events (Salmon, 1984). However, although the explanatory role of science is universally acknowledged, a formal definition of what constitutes and characterises a scientific explanation remains a complex subject. This is attested by the long history of the discussion in Philosophy of Science, which goes back at least to Ancient Greece (Hankinson, 2001). Nevertheless, relatively recent attempts at delivering a rigorous account of scientific explanation have produced a set of quasi-formal models that clarify to some extent the nature of the concept (Salmon, 2006).

The modern view of scientific explanation has its root in the work of Carl Gustav Hempel and Paul Oppenheim, *"Studies in the Logic of Explanation"* (Hempel & Oppenheim, 1948), whose publication in 1948 raised a heated debate in the Philosophy of Science community (Woodward et al., 2017). This section will survey the main accounts resulting from this debate with the aim of summarising and revisiting the main properties of a scientific explanation. In particular, the goal of the survey is to identify the principal constraints that these models impose on *explanatory arguments*, highlighting their essential features and function. This analysis will lead to the comprehension of the essential characteristics that differentiate explanation from other types of knowledge in science, such as prediction.

In general, an explanation can be described as an argument composed of two main elements:

1. The *Explanandum*: the fact representing the observation or event to be explained.
2. The *Explanans*: the set of facts that are invoked and assembled to produce the explanation.

The aim of a formal account of a scientific explanation is to define an *"objective relationship"* that connects the explanandum to the explanans (Salmon, 1984), imposing constraints on the class of possible arguments that constitute a valid explanation. Existing accounts, therefore, can be classified according to the nature of the relationship between explanans and explanandum (Table 1). Specifically, this survey will focus on accounts falling under two main conceptions:

- *Epistemic:* The explanation is an *argument* showing how the explanandum *can be derived* from the explanans. There is a relation of *logical necessity* between the explanatory statements and the event to be explained.

Table 1 The main modern accounts of scientific explanation in Philosophy of Science

Account	Explanans	Relation
Epistemic		
Deductive-Nomological (Hempel & Oppenheim, 1948)	Initial conditions + at least a universal law of nature	The *explanandum* is logically deducedfrom the *explanans*
Inductive-Statistical (Hempel, 1965)	Initial conditions + at least a statistical law	The *explanans* make the *explanandum* highly probable
Unificationist (Kitcher, 1989)	A theory T + a class of phenomena P including the *explanandum*	Shows how a class of phenomena P can be derived from a theory T through the instantiation of an argument pattern
Ontic		
Statistical-Relevance (Salmon, 1971)	A set of statistically relevant facts	the *explanans* increase the probability of the *explanandum*
Causal-Mechanical (Salmon, 1984)	A set of relevant causal processes and interactions	The *explanans* are part of the causal his- tory of the *explanandum*; the *explanans* are part of the mechanism underlying the *explanandum*

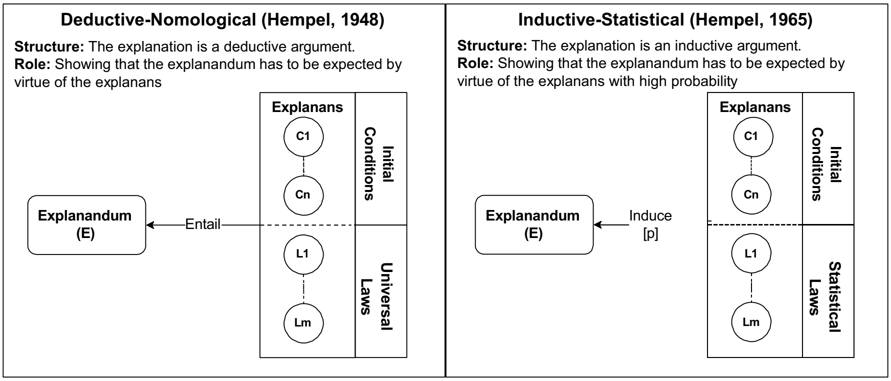

Fig. 1 Schematic representation of the Deductive-Inductive accounts of scientific explanation

- *Ontic:* The explanation relates the explanandum to *antecedent conditions* by means of general laws, *fitting* the explanandum into a *discernible pattern.*

2.1 Explanation as an Argument

2.1.1 Deductive-Inductive Arguments

The *Deductive-Nomological (DN)* model proposed by Hempel (Hempel & Oppenheim, 1948) is consid- ered the first modern attempt to formally characterise the concept of scientific explanation (Fig. 1). the DN account defines an explanation as an argument, connecting explanans and explanandum by means of *logical deduction*. Specifically, the explanantia constitute the premises of a deductive argument while the explanandum represents its logical conclusion. The general structure of the DN model can be schematised as follows:

$C_I, C_2 \ldots, C_k$ Initial Conditions

$L_I, L_2 \ldots, L_r$ Universal Laws of Nature

E. Explanandum

In this model, the explanans constitutes a set of initial conditions, $C_1, C_2 \ldots, C_k$, plus at least a universal law of nature, $L_1, L_2 \ldots, L_r$ (with $r > 0$). According to Hempel, in order to represent a valid scientific argument, an explanation must include only explanans that are empirically testable. At the same time, the universal law must be a statemet describing a *universal* regularity, while the initial conditions represent particular facts or phenomena that are concurrently observable with the explanandum. Here is a concrete example of a scientific explanation under the DN account (Hempel, 1965):

- C_1: The (cool) sample of mercury was placed in hot water;
- C_2: Mercury is a metal;
- L_1: All metals expand when heated;
- E: The sample of mercury expanded.

To complete the DN account with a theory of statistical explanation, Hempel introduced the *Inductive- Statistical (IS)* model (Hempel, 1965). According to the IS account, an explanation must show that the explanandum was to be expected with *high probability* given the explanans. Specifically, an explanation under the IS account has the same structure of the DN account, replacing the universal laws with statistical laws. In order for a statistical explanation to be appropriate, the explanandum must be induced from statistical laws and initial conditions with probability close to 1.

The Deductive-Inductive view proposed by Hempel emphasises the *predictive power* of explanations. Given a universal/statistical law and a set of initial conditions, it is possible to establish whether or not a particular phenomenon will occur in the future. According to Hempel, in fact, explanations and predictions share exactly the *same logical and functional structure*. Specifically, the only difference between explanatory and predictive arguments is when they are formulated or requested: explanations are generally required for past phenomena, while predictions for events that have yet to occur.

This feature of the Deductive-Inductive account is known as the *symmetry thesis* (Hempel, 1965) which has been largely criticised by other philosophers in the field (Kitcher, 1989; Salmon, 1984). The symmetry thesis, in fact, leads to well-known objections and criticisms of Hempel's account. Consider the following example:

- C_1: The elevation of the sun in the sky is x;
- C_2: The height of the flagpole is y;
- L_1: Laws of physics concerning the propagation of light;
- L_2: Geometric laws;
- E: The length of the shadow is z.

While the example above represents a reasonable explanatory argument, the DN account does not impose any constraint that prevents the interchanging of the explanandum with some of the initial conditions:

- C_1: The elevation of the sun in the sky is x;
- C_2: The length of the shadow is z;
- L_1: Laws of physics concerning the propagation of light;
- L_2: Geometric laws;
- E: The height of the flagpole is y.

The DN model and its symmetry property, in particular, allows for the construction of explanatory arguments that contain inverted causal relations between its elements. This counterexample shows that prediction and explanation *must have a different logical structure* and treated as different types of arguments. Although predictive power is a necessary property of an adequate explanation, it is not sufficient. Explanations, in fact, are inherently *asymmetric*, a property that cannot be described by means of deductive-inductive arguments alone.

In Hempel's account, moreover, there is a further property of explanation that has been subject to criticisms by subsequent philosophers, that is the notion of *explanatory relevance*. Consider the following counter-example from Salmon (1984):

- C_1: John Jones is a male;
- C_2: John Jones has been taking birth control pills regularly;
- L_1: Males who take birth control pills regularly fail to get pregnant;
- E: John Jones fails to get pregnant.

Although the argument is formally correct, it contains statements that are explanatorily irrelevant to E. Specifically, the fact that *John Jones has taken birth control pills* should not be cited in an explanation for *John Jones fails to get pregnant*. In this particular example only C_1 is relevant to E, and only C_1 should figure into an explanation for E. Specifically, the universality and high probability requirements of the DN and IS model constrain the explanation to include all the explanatory relevant premises but not to exclude irrelevant facts (Salmon, 1984).

2.1.2 Explanatory Unification and Argument Patterns

The Unificationist account of scientific explanation was proposed by Friedman (Friedman, 1974) and subsequently refined by Kitcher (Kitcher, 1981, 1989) in order to overcome the criticisms, including relevance and asymmetry, raised by the Deductive-Inductive account.

According to the Unificationist model, an explanation cannot be uniquely described in terms of deductive or inductive arguments. To properly characterise an explanation, in fact, it is necessary to consider its main function of fitting the explanandum into a *broader unifying pattern*. Specifically, an explanation is an argument whose role is to connect a set of *apparently unrelated phenomena*, showing that they can be subsumed under a common underlying regularity. The concept of explanatory unification is directly related to the goal of Science of understanding nature by reducing the number of disconnected phenomena and provide an ordered and clear picture of the world (Schurz, 1999).

Figure 2 shows a schematic representation of the Unificationist account. Given a scientific theory T and a class of phenomena P including the explanandum E, an explanation is an argument that allows deriving all the phenomena in P from T. In this case, we say that T *unifies* the explanandum E with the other phenomena in P. According to Kitcher, a scientific explanation accomplishes unification by deriving descriptions of many phenomena through the same patterns of derivation (Kitcher,

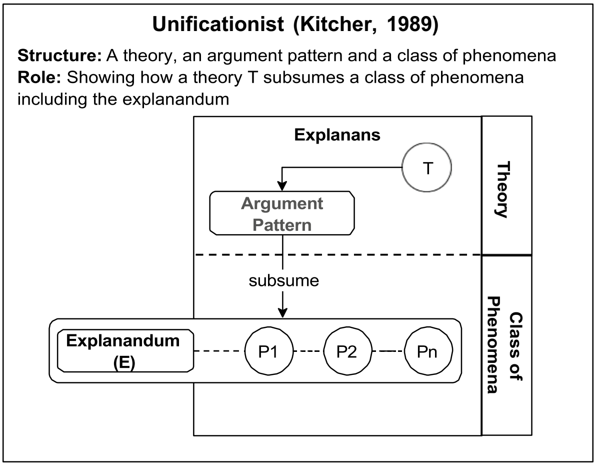

Fig. 2 A schematic representation of the Unificationist account of scientific explanation

1989). Specifically, a theory defines an *argument pattern* which can be occasionally instantiated to explain particular phenomena or observations.

An argument pattern is a sequence of *schematic sentences* organised in premises and conclusions. In particular, a schematic sentence can be described as a template obtained by replacing some non-logical expressions in a sentence with *variables* or *dummy letters*. For instance, from the statement *"Organisms homozygous for the sickling allele develop sickle-cell anemia"* it is possible to generate schematic sentences at different levels of abstraction: *"Organisms homozygous for A develop P"* and *"For all x, if x is O and A then x is P"*. An argument pattern can be instantiated by specifying a set of *filling instructions* for replacing the variables of the schematic sentences together with rules of inference for the derivation. For example, a possible filling instruction for the schematic sentence *"Organisms homozygous for A develop P"* might specify that *A* must be substituted by the name of an allele and *P* by some phenotypic trait. Different theories can induce different argument patterns whose structure is not defined a-priori as in the case of Hempel's account. However, once a theory is accepted, the same argument pattern can be instantiated to explain a large variety of phenomena depending on the unification power of the theory.

The history of science is full of theories and explanations performing unification, and the advancement of science itself can be seen as a process of growing unification (Friedman, 1974). A famous example is provided by Newton's law of universal gravitation, which unifies the motion of celestial bodies and falling objects on Earth showing that they are all manifestation of the same underlying physical law. Specifically, from the unificationist point of view, Newton's law of universal gravitation defines an argument pattern whose filling instructions apply to all entities with mass.

The Unificationist account provides a set of criteria to identify the *"best explanation"* among competing theories:

1. *Unification power*: Given a set of phenomena P and a theory T. the larger the cardinality of P—i.e. the number of phenomena that are unified by T, the greater the explanatory power of T.
2. *Simplicity*: Given two theories T and T_1 able to unify the same set of phenomena P, the theory that makes use of a lower number of premises in its argument patterns is the one with the greatest explanatory power.

These selection criteria play a fundamental role in the Unificationist account since, according to Kitcher, only the best explanation available at a given point in time should be considered as the valid one (Kitcher, 1981). For example, to explain the motion of celestial bodies by means of gravity, one must consider Einstein's theory of relativity as the valid explanation, as it allows to subsume a broader set of phenomena compared to Newton's law of universal gravitation.

The simplicity criteria prevents the explanation to include irrelevant premises as in the case of the control pill example analysed under the Deductive-Inductive account since, under the same unification power, an explanation containing fewer premises will be preferred over a more complex explanation introducing unnecessary statements. Similarly, the problem of asymmetry can be solved considering the unification power criteria. Specifically, argument patterns containing inverted causal relations will generally allow for the derivation of fewer phenomena. According to Kitcher, in fact, causality is an emergent property of unification: *"to explain is to fit the phenomena into a unified picture insofar as we can. What emerges in the limit of this process is nothing less than the causal structure of the world"* (Kitcher, 1989).

2.2 Fitting the Explanandum into a Discernible Pattern

2.2.1 Statistical-Relevance

Motivated by the problem of relevance in the Deductive-Inductive account, Wesley Salmon elaborated a statistical account of explanation known as *Statistical Relevance (SR)* (Salmon, 1971). Differently from the Deductive-Inductive account, the SR model does not concern with the general structure and organisation of the explanatory argument, but attempts to characterise a scientific explanation in terms of the intrinsic relation between each explanatory statement and the explanandum (Fig. 3, left).

In general, given a population A, a factor C and some event B, we say that C is *statistically relevant* to the occurrence of B if and only if

$$P(B|A.C) \neq P(B|A) \text{ or } P(B|A.C) \neq P(B|A.\neg C) \tag{1}$$

In other words, a given factor C is statistically relevant to an event B if its occurrence changes the probability of B to occur. According to the SR account, the explanatory relevance of a fact has to be defined in terms of its statistical relevance. Specifically, an explanation is an *assembly of statistically relevant facts* that increase the probability of the explanandum.

Consider the birth control pills example analysed under the IS account:

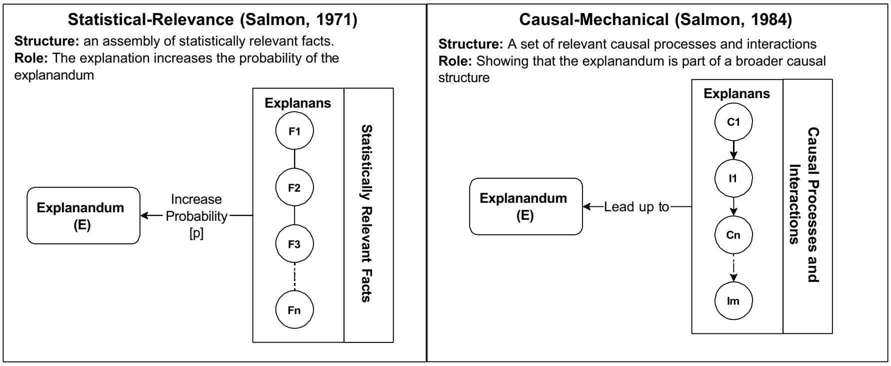

Fig. 3 Schematic representation of accounts falling under the *ontic* conception

1. C_1: John Jones is a male;
2. C_2: John Jones has been taking birth control pills regularly;
3. E: John Jones fails to get pregnant.

Given a population T , we can perform a statistical analysis to verify whether C_1 and C_2 are relevant to E:

$$P(pregnant|T.male) = P(pregnant|T.male.pills) \tag{2}$$

$$P(pregnant|T.pills) \neq P(pregnant|T.pills.male) \tag{3}$$

Notice that in (2.2), given the fact that a generic $x \in T$ is a male (*T.male*), the action of taking birth control pills (*T.male.pills*) has no affect on the probability that x is pregnant. Conversely, in (2.3), the probability that a generic member of the population x is pregnant, given the action of taking pills (*T.pills*), decreases to zero if we know that x is a male (*T.pills.male*). Therefore, the statistical relevance analysis leads to the conclusion that *"among males, taking birth control pills is explanatorily irrelevant to pregnancy, while being male is relevant"* (Salmon, 1984).

The SR model shows that a fact can be explanatorily relevant even if it does not induce the explanandum with a probability close to 1. Specifically, the relevance depends on the effect that the explanans have on the probability of the explanandum rather than on its absolute value. Contrary to the Inductive-Statistical account, this property guarantees the possibility to formulate explanations for rare phenomena.

Although statistical relevance seemed to provide a formal way to shield explanation from irrelevance, Salmon subsequently realised that the SR model is not sufficient to elaborate an adequate account of scientific explanation (Salmon, 1984, 1998). It is nowadays clear, in fact, that certain causal structures are greatly underdetermined by statistical relevance (Pearl, 2009, 2019). Specifically, different causal structures can be described by the same statistical relevance relationships among their elements, making it impossible to discriminate direct causal links by means of statistical relevance analysis alone (Fig. 4).

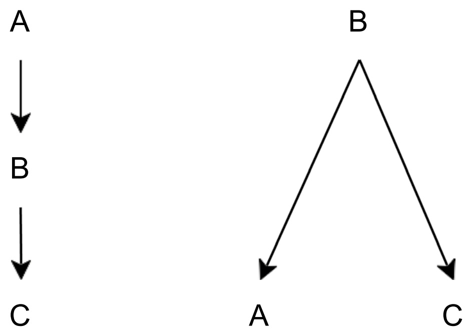

Fig. 4 Causal relationships are underdetermined by statistical relevance relationships. In this example, in particular, it is not possible to discriminate between the depicted causal structures using a statistical relevance analysis. In both cases, in fact, *A* is statistically relevant to *C*; a factor that can lead, in the situation depicted on the right, to a SR explanation based on the relation between *A* and *C* induced by the common cause *B*

According to Salmon, *"the statistical relationships specified in the SR model constitute the statistical basis for a bona-fide scientific explanation, but this basis must be supplemented by certain causal factors in order to constitute a satisfactory scientific explanation"* (Salmon, 1984). The failed attempt to characterise a scientific explanation uniquely in terms of statistical elements demonstrated, as in the case of Hempel's account, the intrinsic difference between prediction and explanation. The latter, in fact, cannot be derived by pure statistical observations and seems to require conjectures and hypotheses about hidden structures, such as the one induced by causal relations and interactions.

2.2.2 Causes and Mechanisms

Following the observation that the SR model is not sufficient for characterising a scientific explanation, Salmon formulated a new account known as the Causal-Mechanical (CM) model, in which the role of an explanation is to show how the explanandum fits into the *causal structure of the world* (Fig. 3, right). Specifically, a valid scientific explanation cannot be limited to statistical relevance and must *cite part of the causal history* leading up to the explanandum.

To formalise the CM account, Salmon attempted to define a theory of causality based on the concepts of *causal processes* and *interactions* (Salmon, 1998). Consider the following example from (Woodward, 2005): *"a cue ball, set in motion by the impact of a cue stick, strikes a stationary 8 ball with the result that the 8 ball is put in motion and the cue ball changes direction"*. Here, the cue ball, the cue stick and the 8 ball are *causal processes* while the collisions between the objects are *causal interactions*. According to the CM model, the motion of the 8 ball has to be explained in terms of the causal processes and interactions belonging to its causal history. Therefore, a generic event *X* is explanatorily relevant to the explanandum *E* if and only if the following conditions apply:

1. *X* is statistically relevant to *E*.
2. *X* and *E* are part of different causal processes.
3. There exists a sequence of causal processes and interactions between X and E leading up to E Salmon identifies two major ways of constructing causal explanations for some event E. An explanation can be either *etiological* – i.e. E is explained by revealing part of its causes – or *constitutive* – i.e. the explanation of E describes the underlying mechanism giving rise to E. A mechanism, in particular, is often described as an organised set of entities and activities, whose interaction is responsible for the emergence of a phenomenon (Craver & Bechtel, 2007; Craver & Tabery, 2015). For example, it is possible to formulate an etiological explanation of a certain disease by appealing to a particular virus, or provide a constitutive explanation describing the underlying mechanisms by which the virus causes the disease.

The foremost merit of the CM account is to exhibit the profound connection between causality, mechanisms, and explanation, highlighting how most of the fundamental characteristics of a scientific explanation derive from its causal nature. Moreover, the differentiation between etiological and constitutive explanation had a significant impact on several scientific fields. Discovering mechanistic explanations, in fact, is nowadays acknowledged as the ultimate goal of many scientific disciplines such as biology and natural sciences (Bechtel & Abrahamsen, 2005; Craver & Darden, 2013; Craver & Tabery, 2015; Schickore, 2014).

The CM model is still subject to a number of criticisms concerning the concepts of causal processes and interactions, which has led subsequent philosophers to propose new theories of causality (Lewis, 1986; Woodward, 2005; Hitchcock, 1995). However, the causal nature of scientific explanations is largely accepted, with much of the contemporary discussion focusing on philosophical and metaphysical aspects concerning causes and effects (Pearl, 2009).

An additional criticism is related to the inherent incompleteness of causal explanations (Craver & Kaplan, 2020; Hesslow, 1988). Since the causes of some event can be traced back indefinitely, causal explanations must show only part of the causal history of the explanandum. This implies that not all the causes of an event can be included in an explanation. In Salmon's account, however, it is not clear what are the criteria that guide the inclusion of relevant causes and the exclusion of others. Subsequent philosophers claimed that the problem of relevance is context-dependent and that it can be only addressed by looking at explanations from a pragmatic perspective (Van Fraassen, 1980). All why questions, in fact, seem to be *contrastive* in nature (Lipton, 1990; Miller, 2018a). Specifically, once a causal model is known, the explanans selected for a particular explanation depend on the specific why question, including only those causes that *make the difference* between the occurrence of the explanandum and some *contrast case* implied by the question (Hilton, 1990; Miller, 2018b) (Table 2).

Table 2 Models of causal attribution adopted to answer different causal questions as defined by varying contrast cases (Hilton, 1990)

Type of implied question	Type of contrast case	Type of cause
"Why X rather than not X?"	Non occurrence of effect	Sum of necessary conditions
"Why X rather than the default value for X?"	The normal case	Abnormal condition
"Why X rather than Y?"	Noncommon effect	Differentiating factor
"Why X rather than what ought to be the case?"	Prescribed or statutory case	Moral or legal fault
"Why X rather than the ideal value for X?"	Ideal case	Design fault or bug

2.3 Summary

This section presented a detailed overview of the main modern accounts of scientific explanation, dis- cussing their properties and limitations.

Despite the fact a number of open questions remain in the Philosophy of Science community, it is possible to draw the following conclusions:

1. **Explanations and predictions have a different structure.** Any attempt to characterise a scientific explanation uniquely in terms of predictive elements has encountered fundamental issues from both an epistemic and an ontic perspective. An explanation, in fact, cannot be entirely characterised in terms of *deductive-inductive arguments* or *statistical relevance* relationships. This is because predictive power, despite being a necessary property of a scientific explanation, is not a sufficient one.
2. **Explanatory arguments create unification.** From an epistemic perspective, the main function of an explanatory argument is to fit the explanandum into a *broader unifying pattern*. Specifically, an explanation must connect a class of *apparently unrelated phenomena*, showing that they can be subsumed under a common underlying regularity. From a linguistic point of view, the unifying power of explanations produces *argument patterns*, whose instantiation can be used to explain a large variety of phenomena through the same patterns of derivation.
3. **Explanations possess an intrinsic causal-mechanistic nature.** From an ontic perspective, a scien- tific explanation cites part of the causal history of the explanandum, fitting the event to be explained into a *causal nexus*. There are two possible ways of constructing causal explanations: (1) an ex- planation can be *etiological* – i.e., the explanandum is explained by revealing part of its causes– or (2) *constitutive* – i.e., the explanation describes the underlying mechanism giving rise to the explanandum.

Philosophers tend to agree that the causal and unificationist accounts are complementary to each other, advocating for a *"peaceful coexistence"* and a pluralistic view of scientific explanation (Bangu, 2017; Glennan, 2002; Salmon, 2006; Strevens, 2004; Woodward et al., 2017). Unification, in fact, seems to be an essential property of causal explanations since many physical processes are the result of the same underlying causal mechanisms (Salmon, 1998, 2006). At the same time, the unifying power of constitutive explanations derives from the existence of mechanisms that have a common higher-level structure, despite differences in the specific entities composing them (Glennan, 2002).

Moreover, the unificationist account seems to be connected with theories of explanation and understand- ing in cognitive science, which emphasise the relationship between the process of searching for broader regularities and patterns to the way humans construct explanations in everyday life through abductive reasoning, abstraction, and analogies (Keil, 2006; Lombrozo, 2006, 2012).

3 Scientific Explanation: The Linguistic Perspective

The previous section focused on the notion of a scientific explanation from a quasi-formal (categorical) perspective, reviewing the main epistemological accounts attempting to characterise the space of explana- tory arguments. Following this survey, this section assumes a linguistic perspective, investigating how the main features of the accepted accounts manifest in *natural language*.

To this end, we present a systematic analysis of corpora of scientific explanations in natural language adopting a mixture of qualitative and quantitative methodologies to investigate the emergence of *explana- tory patterns* at both *sentence* and *inter-sentence* level, relating them to the *Causal-Mechanical* (Salmon, 1998) and *Unificationist* account (Kitcher, 1981, 1989). Specifically, we hypothesise that it is possible to map linguistic aspects emerging in natural language explanations to the discussed models of scientific explanation. At the same time, we observe that some linguistic and pragmatic elements in natural language explanations are not considered by the epistemological accounts, and therefore expect the corpus analysis to provide complementary insights on the nature of explanations as manifested in natural language. Bridging the gap between these two domains aims to provide a necessary linguistic-epistemological grounding for the construction of explanation-based AI models.

The presented analysis focuses on two distinct corpora of explanations; the *Biology Why Corpus* [1] (Jansen et al., 2014), a dataset of biology why-questions with one or more explanatory passages identified in an undergraduate textbook, and the *WorldTree Corpus* [2] (Xie et al., 2020), a corpus of science exams questions curated with natural language explanations supporting the correct answers.

[1] https://allenai.org/data/biology-how-why-corpus

[2] http://cognitiveai.org/explanationbank/

The main features of the selected corpora are summarised in Table 3. As shown in the table, the corpora have complementary characteristics. The explanations included in the *Biology Why Corpus* are specific to a scientific domain (biology in this case), while the *WorldTree Corpus* expresses a more diverse set of topics, including physics, biology, and geology. Since the explanatory passages from the *Biology Why Corpus* are extracted from textbooks, the explanations tend to be more technical and unstructured. On the other hand, the explanations in *WorldTree* are

Table 3 Main features of the analysed corpora of natural language explanations

Feature	Why Corpus	WorldTree
Size	193	2206
Domain	Biology	Science exams
Type	Scientific	Scientific, Commonsense
Annotation	Textbooks	Manually curated c
Structured	No	Yes
Reuse	No	Yes

manually curated and represented in a semi-structured format (aiming more closely on inference automation), often integrating scientific sentences with commonsense knowledge. Moreover, the individual explanatory sentences in *WorldTree* are reused across different science questions when possible, facilitating a quantitative study on knowledge use and the emergence of sentence-level explanatory patterns (Jansen, 2017).

By leveraging the complementary characteristics of the selected corpora and relating the corpus analysis to the discussed accounts of scientific explanation, we aim at investigating the following research questions:

1. **RQ1:** What kinds of explanatory sentences occur in natural language explanations?
2. **RQ2:** How do explanatory patterns emerge in natural language explanations?

We adopt the *Biology Why Corpus* and *WorldTree* to investigate **RQ1**, while *WorldTree* is considered for **RQ2** because of its size and reuse-oriented design.

3.1 Biology Why Questions

To study and investigate the emergence of sentence-level explanatory patterns in biological explanations we performed a systematic annotation of the explanatory passages included in the *Biology Why Corpus* (Jansen et al., 2014). To this end, we identified a set of 11 recurring knowledge categories, annotating a sample of 50 explanations extracted from the corpus. Examples of annotated explanation sentences and their respective knowledge types are included in Table 4.

3.1.1 Recurring Explanatory Sentences

Figure 5 reports the relative frequencies of each knowledge category in the annotated why-questions.

The corpus analysis reveals that the majority of the why questions (75%) are answered through the direct description of *processes* and *mechanisms*. As expected,

Table 4 Explanation sentences in the Biology Why Corpus

Explanandum	Explanans	Knowledge Type
It is important for blood transfusions to not occur between individuals with different blood types	Certain bacteria normally present in the body have epi-topes very similar to the A and B carbohydrates	Analogy, Comparison
It is important for blood transfusions to not occur between individuals with different blood types	By responding to the bacterial epitope similar to the B carbohydrate, a person with type A blood makes antibodies that will react with the type B carbohydrate	Process, Mechanism
It is important for blood transfusions to not occur between individuals with different blood types	Matching compatible blood groups is critical for safe blood transfusions	Requirement, Constraint
Inbreeding does not cause evolution directly	The Hardy–Weinberg is a principle that describes a hypothetical population that is not evolving	Definition
Inbreeding does not cause evolution directly	The gene pool is modified if mutations alter alleles or if entire genes are deleted or duplicated	Conditional, If–then
Inbreeding does not cause evolution directly	Both inbreeding and genetic drift can cause a loss of genetic variation	Causal Interaction
Inbreeding does not cause evolution directly	The allele and genotype frequencies often do change over time	Property, Attribute
Steroids can easily pass through cell membranes	These complexes of a lipid-soluble hormone and its receptor act in the nucleus to regulate transcription of specific genes	Function, Roles
Chromatin is important in meiosis	For example, the nuclei of human somatic cells (all body cells except the reproduc-tive cells) each contain 46 chromosomes	Instances, Examples
It is important for polypeptides to be able to greatly vary in amino acid sequence	Recall that most enzymes are proteins	Taxonimic, Meronymic
Two traits that are more than 50 cM away from each other are inherited randomly relative to each other	The observed frequency of recombination in crosses involving two such genes can have a maximum value of 50%	Probability, Statistical Rel-evance

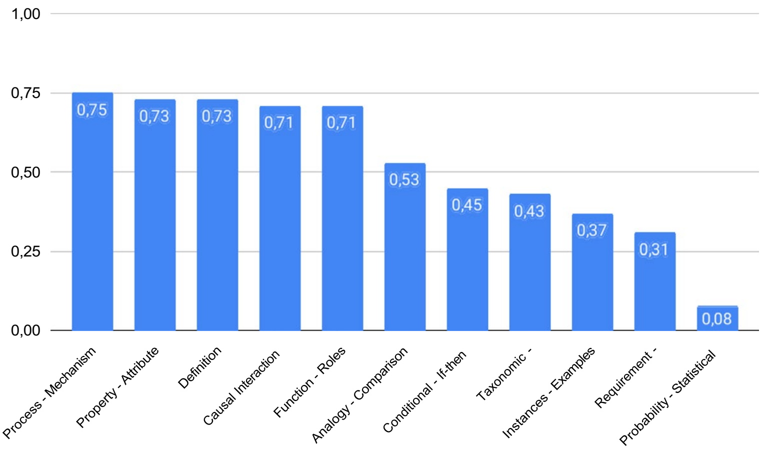

Fig. 5 Recurring knowledge in biological explanations. The graph shows the relative frequencies of different knowledge categories in the annotated Biology Why Corpus

this result confirms the crucial role of *constitutive explanations* as defined in the Causal-Mechanical (CM) account (Salmon, 1984). The importance of causality is confirmed by the frequency of sentences describing direct *causal interactions* between entities (71%), which demonstrates the interplay between *constitutive* and *etiological* explanation. Moreover, the analysis suggests that a large part of the explanations (71%) include sentences describing *functions* and *roles*. The relation between the notion of function and mechanisms is reported in many constitutive accounts of explanation (Craver & Tabery, 2015), and is typically understood as a means of describing and situating some lower-level part within a higher-level mechanism (Craver, 2001).

The corpus analysis suggests that natural language explanations are not limited to causes and mecha- nisms and tend to include additional types of knowledge not explicitly discussed in the epistemological accounts. Specifically, the graph reveals that *definitions* and sentences about *attributes* and *properties* play an equally important role in the explanations (both occurring in 73% of the why questions). We attribute this result to *pragmatic aspects* and inference requirements associated to the *unification* process. Defini- tions, for instance, might serve both as a way to introduce missing context and background knowledge in natural language discourse and, in parallel, as a mechanism for *abstraction*, relating specific terms to high-level conceptual categories (Silva et al., 2018; Silva et al., 2019; Stepanjans & Freitas, 2019).

The role of abstraction in the explanations is supported by the presence of *analogies* and *comparison* between entities (53%), as well as sentences describing *taxonomic* or *meronymic* relations (43%). These characteristics suggest the presence of explanatory arguments performing unification through an abstractive inference process, whose function is to identify common abstract features between concrete instances in the explanandum (Kitcher, 1981). The role of abstraction will be explored in details in the next section.

Table 5 Example of a curated explanation in WorldTree

Explanandum
Two sticks getting warm when rubbed together is an example of a force producing heat
Explanans
1) A stick is a kind of object; (2) To rub together means to move against; (3) Friction is a kind of force; (4) Friction occurs when two object's surfaces move against each other; (5) Friction causes the temperature of an object to increase

Finally, the corpus analysis reveals a low frequency of sentences describing *statistical relevance* rela- tionships and *probabilities* (8%). These results reinforce the fundamental difference between explanatory and predictive arguments identified and discussed in the philosophical accounts (Woodward et al., 2017).

3.2 Science Questions

This section presents a corpus analysis on WorldTree (Xie et al., 2020) aimed at investigating the emergence of explanatory patterns and unification, relating them to epistemological aspects of scientific explanations. Table 5 shows an archetypal example of explanation in WorldTree. Here, the explanandum is represented by a statement derived from a science question and its correct answer, while the explanans are an assembly of sentences retrieved from a background knowledge base.

The corpus categorises the core explanans according to different explanatory roles:

- *Central:* Sentences explaining the central concepts that the question is testing.
- *Grounding:* Sentences linking generic terms in a central sentence with specific instances of those terms in the question.

Some explanatory sentences in WorldTree can be categorised according to additional roles that are not strictly required for the inference (i.e., *Background* and *Lexical Glue* (Jansen et al., 2018)) and that, for the purpose of investigating the nature of explanatory patterns, will not be considered in the corpus analysis.

3.2.1 Distribution and Reuse of Explanatory Sentences

The first analysis concentrates on the distribution and reuse of *central* explanatory sentences in the corpus. The quantitative results of this analysis are presented in Figs. 6 and 7, while a set of qualitative examples are reported in Table 6.

The graph in Fig. 6 describes the distribution of individual sentences annotated as central explanatory facts across different explanations. Specifically, the y-axis represents the number of times a specific sentence is used as a central explanation for a specific science question. The trend in the graph reveals that the occurrence of central explanatory sentences tends to follow a long tail distribution, with a small set of sentences frequently reused across different explanations. This trend suggests that

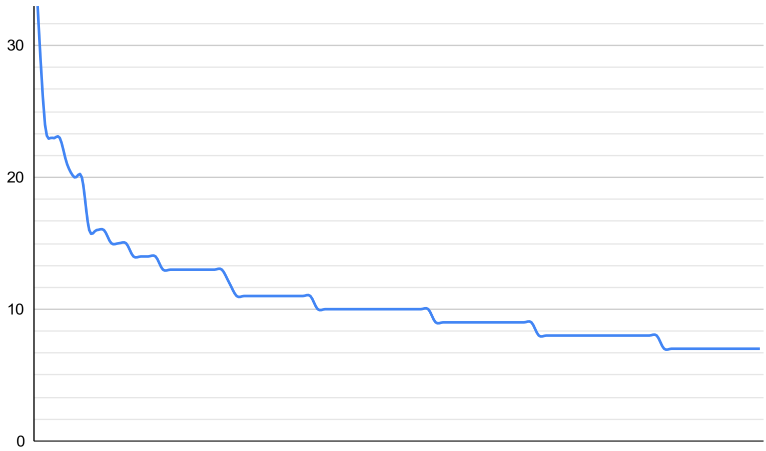

Fig. 6 Distribution and reuse of central explanatory sentences in WorldTree. The y-axis represents the number of times an individual explanatory sentence is used to explain a specific question in the corpus. The trend in the graph reveals that the occurrence of central explanatory sentences tends to follow a long tail distribution, with a small set of sentences frequently reused across different explanations

a subsets of sentences results particularly useful to construct explanations for many science questions, constituting a first indication that some central sentence might possess a greater *explanatory power* and induce certain *patterns of unification*.

To further investigate this aspect, Fig. 7 correlates the frequencies of central explanatory sentences in the corpus (x axis) with the average similarity between the same sentences and the questions they explain (y axis). To perform the analysis, the similarity values are computed adopting BM25 and cosine distance between each question and its explanation sentences (Robertson et al., 2009). From a unificationist point of view, we expect to find an inverse correlation between the frequency of reuse of a central sentence and its similarity with the explanandum. Specifically, we assume that the lower the similarity, the higher the probability that a central sentence describes abstract laws and high-level regularities, and that, therefore, it is able to *unify* a larger set of phenomena. Under these assumptions and considering naturally occurring variability in the dataset, the trend in Fig. 7 confirms the expectation, showing that the most reused central sentences are also the ones that explain clusters of less similar questions. In particular, the graph reinforces the hypothesis that the reuse value of a central sentence in the corpus is indeed connected with its *unification power*.

The concrete examples in Table 6 further support this hypothesis. Specifically, the table shows that it is possible to draw a parallel between the distribution of central sentences in the corpus and the notion of *argument patterns* in the Unificationist account (Kitcher, 1981). It is possible to notice, in fact, that the most occurring central sentences tend to describe high-level processes and regularities, typically mentioning abstract concepts and entities (e.g., *living things*, *object*, *substance*, *material*). In particular, the examples suggest that reoccurring central explanatory facts might act as *schematic sentences* of an *argument pattern*, with abstract entities representing the linguistic counterpart of *variables* and *filling instructions* used to specify and constraining the space of possible instantiations.

3.2.2 Abstraction and Patterns of Unification

To further explore the parallel between natural language explanations and the Unificationist account, we focus on recurring inference chains between *grounding* and *central* sentences. Specifically, we aim to investigate whether it is possible to map inference patterns in WorldTree to the process of instantiating *schematic sentences* for unification. To this end, we automatically build a linkage between grounding and central sentences in the corpus using the support of lexical overlaps.

Table 7 reports the most recurring linguistic categories of grounding-central chains, which provide an indication of the high-level process through which explanatory patterns emerge in natural language. Overall, we found clear evidence of inference patterns related to the *instantiation* of central explanatory sentences. Specifically, the table shows that these patterns emerge through the use of taxonomic knowledge. This suggests that abstraction, intended as the process of going from concrete concepts in the explanandum to high-level concepts in the explanans, is a fundamental part of the inference required for explanation and it is what allows subsuming the explanandum under unifying regularities. Central sentences, in fact, tend to be represented by a more diverse set of linguistic categories in line with those described in the philosophical accounts (i.e., causes, processes, general rules). By looking at grounding-grounding connections one notices the relatively high frequency of chains of taxonomic relations, which confirms again the parallel between explanatory patterns in the corpus and the process of instantiating abstract schematic sentences for unification. Moreover, the presence of linguistic elements about

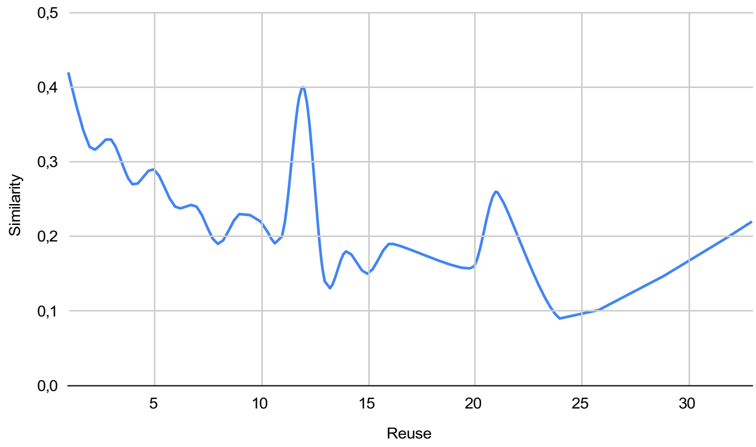

Fig. 7 Similarity between central explanatory sentences and questions against the frequency of reuse of the central explanatory sentences. From a unificationist point of view, we expect an inverse correlation between the frequency of reuse of an explanatory sentence and its similarity with the explanandum. This is because the lower the similarity, the higher the probability that a central sentence describes abstract laws and regularities that *unify* a large set of phenomena

Table 6 Most reused central explanatory sentences in WorldTree

Central Explanatory Sentence	Occurrence
Boiling (evaporation) means matter (a substance) changes from a liquid into a gas by increasing heat energy	33
An adaptation (an ability) has a positive impact on an animal's (living thing's) survival, health, and ability to reproduce	24
Photosynthesis means producers (green plants) convert carbon dioxide, water, and solar energy into carbohydrates, food, and oxygen for themselves	23
Inheriting is when an inherited characteristic is copied and passed from parent to offspring by genetics (DNA)	23
Melting means matter (a substance) changes from a solid into a liquid by increasing heat energy	21
If an object is made of a material then that object has the properties of that material	20
Photosynthesis is a source of food and energy for the plant by converting carbon dioxide, water, and sunlight into carbohydrates	20
Water is in the solid state, called ice, for temperatures between 0, -459, -273 and 273, 32, 0 K, F, C	16
Decomposition is when a decomposer breaks down dead organisms 16 an animal (living things) requires nutrients for survival	16
Objects are made of materials, substances, matter	15
Chemical reactions cause new and different substances to form	15

generic attributes and properties is in line with the analysis on the Biology Why Corpus, supporting the fact that these pragmatic elements in natural language explanations play an important role in the abstraction-instantiation process. Table 8 shows examples of sentence-level explanatory patterns, demonstrating how the process of abstraction and unification concretely manifests in the corpus.

Overall, it is possible to conclude that explanatory patterns emerging in natural language explanations are closely related to unification and that this process is fundamentally supported by an inference mechanism performing abstraction, whose function is to connect the explanandum to the description of high-level patterns and unifying regularities.

3.3 Summary

The main results and findings of the corpus analysis can be summarised as follows:

1. **Natural language explanations are not limited to causes and mechanisms.** While *constitutive* and *etiological* elements represent the core part of an explanation, our analysis suggests that addi- tional knowledge categories such as *definitions*, *properties* and *taxonomic relations* play an equally important role in natural language. This can be attributed to both *pragmatic aspects* of explanations and inference requirements associated to *unification*.
2. **Patterns of unification naturally emerge in corpora of explanations.** Even if not intentionally modelled, *unification* seems to be an emergent property of

Table 7 Most reused categories of grounding-central inference chains in WorldTree

Grounding	Grounding	Occurrence
_ is a kind of _ (Taxonomic)	_ is a kind of _ (Taxonomic)	524
_ is a kind of _ (Taxonomic)	_ is part of _ (Part-of)	73
_ is a kind of _ (Taxonomic)	_ is made of _ (Made-of)	37
_ is a kind of _ (Taxonomic)	_ typically performs action _ on _ (Actions)	30
_ is a kind of _ (Taxonomic)	_ is a property of _ (Properties)	25
Grounding	**Central**	**Occurrence**
_ is a kind of _ (Taxonomic)	_ typically performs action _ on _ (Actions)	209
_ is a kind of _ (Taxonomic)	if _ then _ (Conditionals)	202
_ is a kind of _ (Taxonomic)	_ causes _ (Causal)	179
_ is a kind of _ (Taxonomic)	_ changes from _ to _ by _ (Processses)	153
_ is a kind of _ (Taxonomic)	_ uses _ for _ (Functional)	133

corpora of natural language explanations. The corpus analysis, in fact, reveals that the frequency of reuse of certain explanatory sentences is connected with the notion of *unification power*. Moreover, a qualitative analysis suggests that reused explanatory facts might act as *schematic sentences*, with abstract entities representing the linguistic counterpart of *variables* and *filling instructions* in the Unificationist account.

3. **Unification is realised through a process of abstraction.** Specifically, abstraction represents the fundamental inference mechanism supporting unification in natural language. The corpus analysis, in fact, suggests that recurring explanatory patterns emerge through inference chains connecting concrete instances in the explanandum to high-level concepts in the central explanans. This process, realised through specific linguistic elements such as *definitions* and *taxonomic relations*, is a fundamental part of natural language explanations that subsumes the event to be explained under high-level patterns and unifying regularities.

4 Synthesis

Finally, with the help of Fig. 8, it is possible to perform a synthesis between the epistemological accounts of scientific explanation and the linguistic aspects emerging from the corpus analysis.

In general, explanations cannot be exclusively characterised in terms of *inductive* or *deductive* arguments. This is because the logical structure of explanations and predictions is intrinsically different (Woodward et al., 2017). From an epistemic perspective, in fact, the main function of an explanatory argument is to fit the explanandum into a broader pattern that maximises unification, showing that a set of apparently unrelated phenomena are part of a common regularity (Kitcher, 1981, 1989). From a linguistic point of view, the process of unification tends to generate sentence-level *explanatory patterns* that can be

Table 8 Most reused sentence-level inference chains in WorldTree

Grounding	**Grounding**	**Occurrence**
An animal is a kind of living thing	A living thing is a kind of object	18
An animal is a kind of organism	A plant is a kind of organism	14
A human is a kind of animal	An animal is a kind of organism	14
A tree is a kind of plant	A plant is a kind of organism	11
A human is a kind of animal	An animal is a kind of living thing	11
Grounding	**Central**	**Occurrence**
Water is a kind of liquid at room temperature	Boiling;evaporation means matter; a substance changes from a liquid into a gas by increasing heat energy	20
Metal is a kind of material	If an object is made of a material then that object has the properties of that material	14
Earth is a kind of planet	A planet rotating causes cycles of day and night on that planet	9
A plant is a kind of organism	Decomposition is when a decomposer breaks down dead organisms	9
Water is a kind of liquid at room temperature	Freezing means matter; a substance changes from a liquid into a solid by decreasing heat energy	9
Metal is a kind of material	Metal is a thermal; thermal energy conductor	9

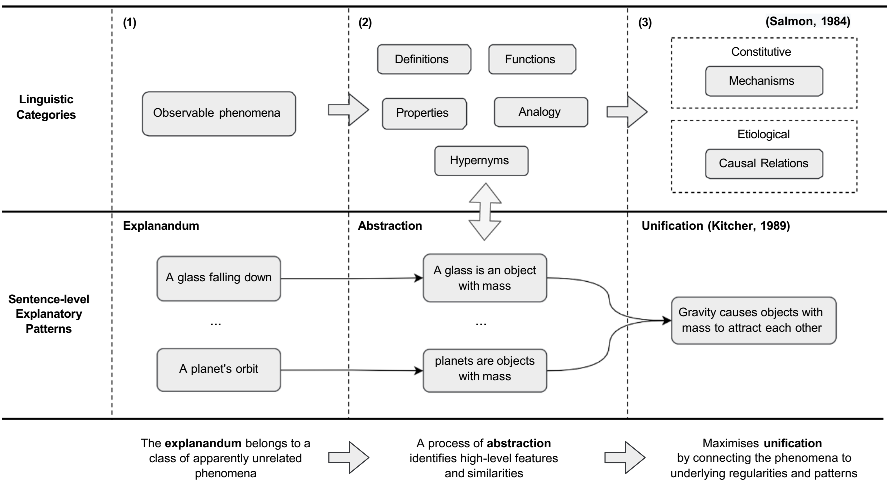

Fig. 8 A synthesis between the formal accounts of scientific explanation and linguistic aspects found through the corpus analysis

reused and instantiated for deriving and explaining many phenomena. In natural language, unification generally emerges as a process of *abstraction* from the explanandum through the implicit search of common high-level features and similarities between different phenomena.

From an ontic perspective, causal interactions and mechanisms constitute the central part of an ex- planation as they make the difference between the occurrence and non-occurrence of the explanandum (Lipton, 1990; Salmon, 1984). Moreover, causal interactions are responsible for high-level regularities and invariants, with many phenomena being the result of the same underlying causal mechanisms. Here, abstraction represents the inference mechanism linking the explanandum to these regularities, a process that manifests in natural language through the use of specific linguistic elements coupled with causes and mechanisms, such as definitions, taxonomic relations, and analogies.

5 Implications for Explanation-based AI

Current lines of research in NLI focus on the development and evaluation of explanation- based models, capable of performing inference through the interpretation and generation of natural language explanations (Jansen et al., 2018; Thayaparan et al., 2021b; Wiegreffe & Marasovic, 2021; Xie et al., 2020). In the context of NLI, explanations aim to support the fundamental goal of improving the applicability of AI models in real-world and high-risk scenarios, enhancing the transparency of the decision-making process for the end

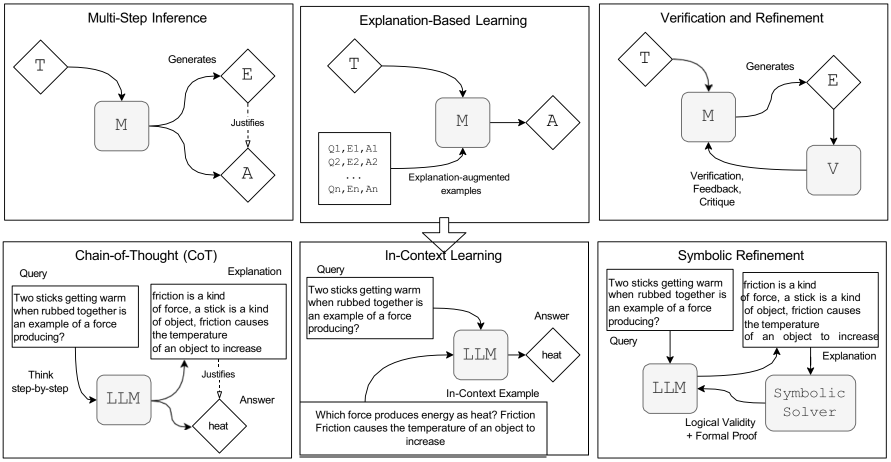

Fig. 9 (Top) a schematic representation of different paradigms under which an explanation-based NLI model (*M*) can leverage explanations (*E*) to produce an answer (*A*) for a given task (*T*). (Bottom) an example of implementation of each paradigm via Large Language Models (LLMs) in the context of question-answering and natural language inference tasks

user, as well as the controllability, alignment, and intrinsic reasoning capabilities of the models.
Given a generic task T , an explanation E can be integrated in an AI model M using different paradigms to generate an answer A for the task T (Fig. 9, Top):

1. **Multi-Step Inference:** The model M can be explicitly designed and trained to generate a step-by-step explanation E to derive the answer A for a task T. In this case, the explanation acts as a justification for the generated output, potentially improving transparency and the ability to break down complex tasks into multiple, sequential reasoning steps (Wei et al., 2022; Yao et al., 2024).
2. **Explanation-Based Learning:** The explanation E is adopted as an additional training signal for a model M. In this context, traditional training sets consisting of input–output pairs are augmented with human-annotated or synthetically generated explanations that describe the explicit reasoning required to solve instances of T (Thayaparan et al., 2020; Wiegreffe & Marasovic, 2021). The additional training signal provided by the explanations can act as a demonstration to improve the generalisation of the model M on unseen problems and make the training process more efficient.
3. **Verification and Refinement:** The explanation E generated by a model M can be used by an external system V to evaluate the quality of the output generated by M. In turn, V can produce detailed feedback and critiques for refining the output of M and provide a formal or empirical assessment of its behaviour (Madaan et al., 2024; Quan et al., 2024).

While these explanation-based paradigms can be instantiated with different architectures in different domains, they are becoming widespread in the subfield of Natural Language Processing (NLP), where recent progress supported by Large Language Models (LLMs) (Kojima et al., 2022; Vaswani et al., 2017) has enabled the automatic processing and generation of explanatory arguments at scale. In this context, the multi-step inference paradigm is typically realised via specific prompting techniques (Qiao et al., 2023) (e.g., Chain-of-Thoughts (Wei et al., 2022), Tree-of-Thoughts (Yao et al., 2024)) where an LLM can be trained and prompted to generate step-by-step explanations to solve specific tasks (e.g., question-answering, natural language inference). Similarly, Explanation-Based Learning is typically realised through a technique known as In-Context Learning (Dong et al., 2022), where demonstration examples and their solutions are provided to the model to guide the generation of answers for unseen problems. Moreover, the generative capabilities of LLMs have enabled the implementation of verification and refinement methods (Gu et al., 2023; Ling et al., 2024; Madaan et al., 2024) often with the help of external tools (Schick et al., 2024), where LLMs' explanations can be first translated into a formal language (e.g., first-order logic) and then verified through symbolic solvers that can provide detailed feedback in the form of logical proofs for subsequent improvements (Dalal et al., 2024; Quan et al., 2024). However, while explanations play a crucial role in state-of-the-art AI models, existing evaluation frameworks for assessing the quality and properties of natural language explanations are still limited (Jansen et al., 2021; Valentino et al., 2021a). Most of the existing evaluation methods, in fact, focus on unidimensional inferential properties defined in terms of *entailment* relationships between explanation and predicted answer (Camburu et al., 2018; Dalvi et al., 2021; Valentino et al., 2021a; Yang et al., 2018). However, our analysis shows that natural language explanations cannot be reduced exclusively to deductive reasoning or entailment relationships. This is because deductive arguments cannot fully characterise explanations, and cannot distinguish explanatory arguments from mere predictive ones. As the function of an explanation includes performing abstraction and unification through recurring explanatory patterns, the evaluation methodologies should move from unidimensional evaluation metrics to multidimensional ones (Dalal et al., 2024), considering diverse linguistic and logical features. Moreover, a more advanced theoretical awareness of explanatory properties could help AI scientists formulate clearer hypotheses to understand the strengths and limitations of explanation-based methods such as Chain-of-Thought and In-Context Learning, whose inferential mechanisms are still largely unknown and debated in the research community (Min et al., 2022; Turpin et al., 2024).

Emergent unification patterns in natural language explanations, for example, have been shown to provide a way to build more robust and efficient multi-step inference models (Valentino et al., 2021b). Similarly, recurring explanatory patterns in pre-trained corpora or in-context examples could help explain the behaviour of explanation-based methods for LLMs in terms of reduced search space deriving from patterns of abstraction and unification (Erasmus & Brunet, 2022; Thayaparan et al., 2021a; Valentino et al., 2022a, 2022b; Zheng et al., 2023).

Regarding the construction of explanation-augmented datasets, while we show that unification seems to be an emergent property of existing corpora (Jansen et al., 2018; Xie et al., 2020), future research can benefit from explicitly considering the presented theoretical and linguistic analysis when designing the explanation annotation process. Unification patterns, for example, can provide a top-down and schema-oriented methodology to scale up the annotation process and help assess a multi-dimensional set of properties including abstraction, the identification of underlying invariants and causal mechanisms as well as the ability to consistently connect multiple instances of the same problem under unifying high-level explanatory regularities.

6 Conclusion

In order to provide an epistemologically grounded characterisation of natural language explanations, this paper attempted to bridge the gap between theory and practice on the notion of *scientific explanation* (Salmon, 1984, 2006), studying it as a *formal object* and a *linguistic expression*. The combination of a systematic survey with a corpus analysis on natural language explanations (Jansen et al., 2014, 2018) allowed us to derive specific conclusions on the nature of explanatory arguments from both a top-down (categorical) and a bottom-up (corpus-based) perspective:

1. Explanations cannot be entirely characterised in terms of *inductive* or *deductive* arguments as their main function is to perform *unification*.
2. A scientific explanation typically cites causes and mechanisms that are responsible for the occurrence of the explanandum.
3. While natural language explanations possess an intrinsic causal-mechanistic nature, they are not limited to causes and mechanisms.
4. Patterns of unification naturally emerge in corpora of explanations even if not intentionally modelled.
5. Unification emerges through a process of abstraction, whose function is to provide the inference mechanism for subsuming the event to be explained under recurring patterns and regularities.

From these findings, it is possible to derive a set of guidilines for furure research on NLI for the creation and evaluation of models that can interpret and generate natural language explanations:

1. Explanations generated by AI models cannot be evaluated only in terms of deductive inference capa- bilities and entailment properties. This is because deductive arguments cannot entirely characterise explanations, and cannot be used to distinguish explanatory arguments from mere predictive ones.
2. As the main function of an explanatory argument is to perform unification, the evaluation methodolo- gies should explicitly consider such property. Moreover, while unification seems to be an emergent property of existing benchmarks, future work might benefit from building a top-down, schema- oriented approach for the creation of explanation-augmented corpora to facilitate evaluation.

3. From a bottom-up perspective, the evaluation of explanations should move from unidimensional met- rics to a multidimensional perspective analysing multiple linguistic and logical properties, including causality, abstraction, the interpretation of definitions, and the ability to make analogies between apparently different problems.
4. The unification property of explanatory arguments can provide a way to build more robust inference models, as well as more efficient and scalable solutions to construct explanation-augmented corpora. Moreover, the emergence of recurring explanatory patterns in large corpora and in-context examples can help explain the success of recent explanation-based methods (e.g., Chain-of-Thoughts, In-Context Learning) as they can reduce the search space for multi-step inference models and support a more schematic, reuse-oriented mechanism for inference on unseen test examples.

The paper contributed to addressing a fundamental gap in classical theoretical accounts on the nature of scientific explanations and their materialisation as linguistic artefacts, providing a unified epistemological- linguistic perspective. We hope such characterisation can support a more principled design and evaluation of explanation-based AI systems which can better interpret and generate natural language explanations.

Acknowledgements This work was partially funded by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617).

Authors' Contributions All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Marco Valentino.

Funding This work was partially funded by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617).

Data Availability The material and code adopted for the corpus analysis is available online at the following URL: https://github.com/ai-systems/scientific_explanations_analysis.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests Not Applicable.

References

Bangu, S. (2017). Scientific explanation and understanding: Unificationism reconsidered. *European Journal for Philosophy of Science, 7*(1), 103–126.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Phi- Losophy of Science Part c: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 421–441.

Cabrera, F. (2021). The fate of explanatory reasoning in the age of big data. *Philosophy & Technology, 34*, 645–665.
Camburu, O. M., Rocktäschel, T., Lukasiewicz, T., & Blunsom, P. (2018). e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa- Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 9539–9549). Curran Associates, Inc.
Camburu, O. M., Shillingford, B., Minervini, P., Lukasiewicz, T., & Blunsom, P. (2020). Make up your mind! adversarial generation of inconsistent natural language explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4157–4165). Online, July. Association for Computational Linguistics. Online, July.
Clinciu, M. A., Eshghi, A., & Hastie, H. (2021). A study of automatic metrics for the evalua- tion of natural language explanations. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2376–2387). Online, April. Association for Computational Linguistics.
Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science, 68*(1), 53–74.
Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology & Philosophy, 22*(4), 547–563.
Craver, C. F., & Kaplan, D. M. (2020). Are more details better? on the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319.
Crave, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.
Craver, C., & Tabery, J. (2015). *Mechanisms in science*.
Dalal, D., Valentino, M., Freitas, A., & Buitelaar, P. (2024). Inference to the best explanation in large language models. *arXiv preprint* arXiv:2402.10767.
Dalvi, B., Jansen, P., Tafjord, O., Xie, Z., Smith, H., Pipatanangkura, L., & Clark, P. (2021). Explaining answers with entailment trees. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7358–7370)
Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable ai for natural language processing. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 447–459)
Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., & Sui, Z. (2022). *A survey on in-context learning*.
Došilović, F. K., Brčić, M., & Hlupić., N. (2018). Explainable artificial intelligence: A survey. *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE.
Erasmus, A., & Brunet, T. D. P. (2022). Interpretability and unification. *Philosophy & Technology, 35*(2), 42.
Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy, 71*(1), 5–19.
Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*(S3), S342–S353.
Gu, Y., Tafjord, O., & Clark, P. (2023). Digital socrates: Evaluating llms through explanation critiques. *arXiv preprint* arXiv:2311.09613.
Hankinson, R. J. (2001). *Cause and explanation in ancient Greek thought*. Oxford University Press.
Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135–175.
Hempel, C. G. (1965). *Aspects of scientific explanation.*
Hesslow, G. (1988). *The problem of causal selection* (pp. 11–32). Contemporary science and natural explanation: Commonsense conceptions of causality.
Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*(1), 65.
Hitchcock, C. R. (1995). Salmon on explanatory relevance. *Philosophy of Science, 62*(2), 304–320.
Jansen, P., Surdeanu, M., & Clark, P. (2014). Discourse complements lexical semantics for non-factoid answer reranking. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 977–986)
Jansen, P., Balasubramanian, N., Surdeanu, M., & Clark, P. (2016). What's in an explanation? char- acterizing knowledge and inference requirements for elementary science exams. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2956–2965)

Jansen, P., Wainwright, S., Marmorstein, S., & Morrison, C. (2018). WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan, May. European Language Resources Association (ELRA).

Jansen, P., Smith, K. J., Moreno, D., & Ortiz, H. (2021). On the challenges of evaluating composi- tional explanations in multi-hop inference: Relevance, completeness, and expert ratings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7529–7542)

Jansen, P. A. (2017). A study of automatically acquiring explanatory inference patterns from corpora of explana- tions: Lessons from elementary science exams. *6th Workshop on Automated Knowledge Base Construction (AKBC 2017).*

Jhamtani, H., & Clark, P. (2020). Learning to explain: Datasets and models for identifying valid rea- son- ing chains in multihop question-answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 137–150). Online, November. Association for Computational Linguistics.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57*, 227–254.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*(4), 507–531.

Kitcher, P. (1989). *Explanatory unification and the causal structure of the world.*

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213.

Lewis, D. (1986). Causal explanation.

Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., & Su, H. (2024). Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems* (p. 36)

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement, 27*, 247–266.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464–470.

Lombrozo, T. (2012). Explanation and abductive inferenc. *Oxford handbook of thinking and reasoning* (pp. 260–276)

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. (2024). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* (p. 36)

Miller, T. (2018a). Contrastive explanation: A structural-model approach. *arXiv preprint* arXiv:1811.03163.

Miller, T. (2018b). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence.*

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 11048–11064)

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*(3), 54–60.

Prasetya, Y. (2022). Anns and unifying explanations: Reply to erasmus, brunet, and fisher. *Philosophy & Technology, 35*(2), 43.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., & Chen, H. (2023). Reasoning with language model prompting: A survey. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5368–5393)

Quan, X., Valentino, M., Dennis, L., & Freitas, A. (2024). Enhancing ethical explanations of large language models through iterative symbolic refinement. In Yvette Graham & Matthew Purver (Eds.), *Pro- ceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1–22). St. Julian's: Association for Computational Linguistics.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press.

Salmon, W. C. (1998). *Causality and explanation*. Oxford University Press.

Salmon, W. C. (1971). *Statistical explanation and statistical relevance, volume 69*. University of Pittsburgh Pre.

Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh press.

Dos Santos Silva, S., Handschuh, S., & Freitas, A. (2018). Recognizing and justifying text entail- ment through distributional navigation on definition graphs. *AAAI* (pp. 4913–4920)

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2024). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* (p. 36)

Schickore, J. (2014). *Scientific discovery*.

Schurz, G. (1999). Explanation as unification. *Synthese*, 95–114.

Silva, V. S., Freitas, A., & Handschuh, S. (2019). Exploring knowledge graphs in an interpretable composite approach for text entailment. *In Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 7023–7030.

Stepanjans, A., & Freitas, A. (2019). Identifying and explaining discriminative attributes. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4313–4322)

Strevens, M. (2004). The causal and unification approaches to explanation unified—causally. *Noûs, 38*(1), 154–176.

Tan, C. (2021). On the diversity and limits of human explanations. *arXiv preprint* arXiv:2106.11988.

Thayaparan, M., Valentino M., & Freitas, A. (2020). A survey on explainability in machine reading comprehension. *arXiv preprint* arXiv:2010.00389.

Thayaparan, M., Valentino, M. & Freitas, A. (2021a). Explainable inference over grounding- abstract chains for science questions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1–12) Online, August. Association for Computational Linguistics.

Thayaparan M., Valentino, M., Jansen, P ., & Ustalov, D. (2021b). TextGraphs 2021 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)* (pp. 156–165), Mexico City, Mexico, June. Association for Computational Linguistics.

Turpin, M., Michael, J., Perez, E., & Bowman, S. (2024). Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* (p. 36)

Valentino, M., Thayaparan, M., Ferreira, D., & Freitas, A. (2022a). Hybrid autoregressive inference for scalable multi-hop explanation regeneration. *In Proceedings of the AAAI Conference on Artificial Intelligence, 36*, 11403–11411.

Valentino, M., Pratt-Hartmann, I., & Freitas, A. (2021a). Do natural language explanations represent valid logical arguments? verifying entailment in explainable nli gold standards. *Proceedings of the 14th International Conference on Computational Semantics (IWCS)* (pp. 76–86)

Valentino, M., Thayaparan, M., & Freitas, A. (2021b). Unification-based reconstruction of multi-hop explanations for science questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 200–211, Online, April. Association for Computational Linguistics.

Valentino, M., Thayaparan, M., & Freitas, A. (2022b). Case-based abductive natural language inference. *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 1556–1568)

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł, & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 5998–6008)

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Wiegreffe, S., & Marasovic, A. (2021). Teach me to explain: A review of datasets for explainable natural language processing. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Woodward, J., Zalta, E. N., et al. (2017). Scientific explanation. *The Stanford*.

Xie, Z., Thiem, S., Martin, J., Wainwright E., Marmorstein, S., & Jansen, P. (2020). WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 5456– 5473). Marseille, France, May. European Language Resources Association.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2369–2380)

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems* (p. 36)

Zheng, H. S., Mishra, S., Chen, X., Cheng H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (2023). Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint* arXiv:2310.06117.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Marco Valentino[1] · **André Freitas**[1,2]

✉ Marco Valentino
marco.valentino@idiap.ch

1 Idiap Research Institute, Martigny, Switzerland

2 University of Manchester, Manchester, United Kingdom