

# A Logic of Best Explanations

April 17, 2017

## 1 Introduction

The past two decades have witnessed ever more sophisticated attempts to formalize explanatory reasoning. The growing literature of logic-based approaches to abduction exhibits both a diversity of perspectives (structural, adaptive, tableau-based, etc.) as well as a panoply of targets—there is not one, but rather many *patterns* of abduction. The most fundamental distinction among these patterns is between the process of generating hypotheses—i.e. *inference to a plausible explanation* (IPE)—and that of evaluating hypotheses—i.e. *inference to the best explanation* (IBE). Unfortunately, work on both patterns faces two significant limitations.

While IPE has received more attention than IBE, there has been little effort among logicians to represent the uniquely *explanatory* character of the premises. In particular, no systematic attempt has been made to formally capture the behavior of the connective(s) expressed by the English locution “That...explains why...” and its adjectivally-modified cousins. Instead, abduction is all too often treated as a kind of ‘deduction in reverse’ with material implication standing in for the explanatory connective.

Logicians’ hesitancy to formalize the explanatory relations expressed in natural language is not surprising; these relations exhibit a Gordian knot of logic properties. They appear to be contradiction-intolerant, (possibly) non-transitive, (probably) asymmetric, (definitely) non-monotonic, and (worst of all) irreflexive. Some of these properties, such as non-transitivity and non-monotonicity are captured by well-established nonclassical logics, but others, such as irreflexivity, run counter to the most liberal conceptions of logical consequence. In an effort to realize some of these properties—and thus to block reflexive, contradictory, or overly-informative explanations—extant abductive logics resort to various *ad hoc* constraints on the application of ‘backwards Modus Ponens’.<sup>1</sup>

Naturally, the problem of accurately representing explanatory relations also plagues logical treatments of IBE. Except, in the case of IBE, this difficulty is compounded by the fact that what constitutes the ‘best explanation’ remains a highly controversial matter, even in philosophy of science where this question receives the most attention. Furthermore, many philosophers of science believe that our best explanations are determined, in part, by contextual considerations. Consequently, the formalization of IBE will need to reserve a parameter for extra-logical criteria. There have been some attempts to meet this challenge by incorporating structured preference rankings, but such an approach treats that the evaluation of

---

<sup>1</sup>A notable exception to this trend is the work of adaptive logicians. In what is perhaps the most successful attempt to do justice to the explanatory character of abductive premises, Mathieu Beirlaen and Atocha Aliseda employ an operator from an ultra-weak conditional logic. The conditional they use, that of Brian Chellas’s **CR**, blocks the validity of *idempotence*, *ex falso quod libet*, *strengthening the antecedent*, *modus ponens*, *contraposition*, and *hypothetical syllogism*—to name just a few. The result is an adaptive logic for (plain) abduction that operates on conditionals that behave in many ways like explanatory expressions do in English. There is, however, no introduction rule for their explanatory conditional—i.e. there is no sense in which one might reason *to* an explanatory proposition.

explanantia as a given rather than as a product of reasoning. At the very least, the apparent context-sensitivity of best explanations presents yet another obstacle to the logical study of IBE.

There are several reasons to persevere in the face of these challenges. First, abandoning the logical study of IBE effectively cedes that task to quantitative approaches while depriving qualitative logicians of a potentially fruitful opportunity to develop their frameworks. Second, when logicians avoid the peculiarities of explanatory relations, philosophers of science lose an invaluable resource, namely the insights that only the construction and manipulation of logical systems can yield. Finally, this abdication severs the link between logical accounts of abduction and the tradition, begun by Carl Hempel, of treating explanations as a species of arguments—a view that we call Explanation-as-Inference (EAI). Over the decades, Hempel’s Deductive-Nomological model of explanation has been subject to withering critiques; but it ought to be remembered that he developed this model long before the advent of sophisticated nonclassical logics. What would the Hempelian model look like were it to be cultivated in the rich soil of today’s nonmonotonic, substructural, or paraconsistent logics? So long as logicians shy away from explanatory relations, we will never know.

This paper attempts to treat these two problems confronting the logic of abduction: the perspicuous representation of explanatory relations and the qualitative formalization of IBE. The key to tackling both is the idea that explanatory relations should neither be reduced to classical deductions nor treated as primitives. Proponents of IBE in philosophy of science (a.k.a explanationists) have long held that a prior grip on explanation can underwrite good non-deductive inference. On our view, this gets things fundamentally backwards: *contra* explanationism, inquirers first engage in non-deductive inference-making, and only late in the game do they baptize some of these inferences as explanations.

Our proposal is to treat the explanatory connective in IBE as an object-language expression for what, in the meta-language, is a rule of defeasible inference. In other words, the *best-explains-why* operator codifies a defeasible consequence relation just as material implication codifies the classical consequence relation. Rather than appeal to a pre-established preference ranking, we use a model of defeasible inference that has two parameters for extra-logical information: one for the set of situations in which an inference would fail and another for information on the actual situation in which the inference is made. By formalizing operations on these two parameters, we are able to provide a uniform proof theory that captures both the structure of IBE as well as the least controversial properties of the *best-explains* connective.<sup>2</sup>

This formal project advances hand-in-hand with a substantive theory of explanation that belongs to the EAI family. Recent work on this theory, what we call the *the defeasibility model of explanation*, has sought to clear the ground for a renewal of the EAI view, to defend the model from criticisms that scuttled previous versions of EAI (e.g. the notorious ‘symmetry problem’), and to situate our approach within a broader theory of logical vocabulary (i.e. inferentialist-expressivism). The present work has a narrower target in view—to present the details of our logic for IBE, to demonstrate that its explanatory connective encodes (a class of) consequence relations that have a wide array of properties associated with “best explains why”, and to show that it preserves the prized logical property of cut-elimination.

<sup>JM</sup>[Road Map](#)

---

<sup>2</sup>We can approach the formalization of explanatory relations with more confidence than our predecessors due to the progress that has been made by logicians investigating the properties of metaphysical explanations in what has come to be known as the *logic of grounding*. In particular, a number of logics have been proposed for the relation of full strict immediate ground, which is widely recognized to be irreflexive, asymmetric, non-transitive, and (in some sense) minimal. While we remain suspicious regarding the extent to which this relation is genuinely *explanatory*, we have been inspired by the innovative work done by logicians in this area research.

## 2 Explanatory Vocabulary

To start, let us specify our target. We propose a logical system whose object-language contains a connective that behaves in a manner similar to a class of natural language locutions. The class we have in mind includes those expressions which are given in English as “That...best explains why...,” and its nominalization “That...is the best explanation of why...,” where the ellipses are filled by declarative sentences. As with most complex locutions in natural language, the analysis of these expressions faces challenges ranging from issues of polysemy to the “non-intersective” behavior of the adjectival predicates. Developing a logic attentive to the various subtleties of these natural language expressions is well beyond the scope of this essay. Instead, we aim to capture at least part of the notion of *best explanation*. Unfortunately, even here, there is ambiguity. For instance, “*A* explains *B*” is typically factive. On a simple reading, then “*A* best explains *B*” is also factive. After all, it is natural to think of our best explanations as a proper subset of the totality of actual explanations. However, our chief aim is to capture the meaning of “*A* best explains *B*” as it functions as a major premise in IBE. In its simplest rendition, this inference takes the following form: *A best explains why B, B / So, A*.

As IBE’s foremost defender, Peter Lipton argues that to treat “*A* best explains *B*” as factive or “actual” in this case is “like a dessert recipe that says start with a soufflé.” (Lipton 2004, 58). For this reason, Lipton suggests that the best explanation ought to be construed as the best *potential* explanation.<sup>3</sup> Following a common convention in the literature on scientific explanation, let us stipulate that “*A* potentially explains *B*” requires neither “*A*” nor “*B*” to be true. Then our claim is that “*A* best explains *B*” is not factive, but must still provide good inductive grounds for detaching the explanans, *A*, when the explanandum, *B*, obtains.

In addition to being non-factive, the sense of *best explanation* that we aim to represent is *immediate* and *exhaustive*. We say that *A* is an immediate explanation of *B* so long as it does not explain *B* merely by explaining something else, *C*, which in turn explains *B*. In other words, the explanations we have in mind are not transitive. By “exhaustive,” we mean that nothing needs to be added to *A* in order for it to explain *B*. The target of our account is thus expressed by the locution “...is the loveliest potential, immediate, exhaustive explanation of why ...”. To avoid having to repeat this phrase, we will henceforth speak of *best explanations* or simply *explanations*. The connective intended to represent this expression is the *best-explains-why* operator or BEW-operator; it is denoted in our formal language by the symbol ‘ $\triangleright$ ’.

## 3 Properties of Explanatory Arguments

As noted, we follow the EAI approach to explanation pioneered by Hempel. In general, EAI holds that explanations are a type of argument or inference. Of course, we are interested in the type of inference that can be identified with our *best* explanations. We will use the term ‘*explanatory argument*’ and the modified turnstile,  $\vdash$ , to refer to this, as-yet-unspecified, consequence relation. Since our aim is to have our BEW-operator exhibit the properties of best explanations, and since we identify the latter with explanatory arguments, in order to make progress in this aim, we must specify the properties of explanatory arguments that distinguish them from other inferences.

<sup>3</sup>Lipton also argues that IBE should construe the best explanation as the “loveliest” explanation, and then vaguely describes “loveliness” as a combination of various “theoretical virtues,” such as simplicity, scope, fit with background belief, etc. While our view does appeal to some superlative—what we call “sturdiness”—it appears to be something quite different than this. A detailed comparison of loveliness and sturdiness exceeds the scope of the current paper.

Explanatory arguments exhibit six distinctive properties. First, they are *defeasible*—an argument can cease to be explanatory when new information is added to its premises. For instance, while the claim that “The liquid’s acidity explains why the blue litmus paper turned red,” may constitute a good explanation, strengthening the explanans can easily produce a bad one: “The liquid’s acidity and the presence of chlorine gas explains why the blue litmus paper turned red.” In this case, the additional information is incompatible with the explanandum. In other cases, however, what is added to an explanans undercuts the explanatory relation itself. For instance, that the match was struck may explain why the match lit, but if it is discovered that match was damp, then the striking no longer, by itself, explains why it lit (perhaps the match was dried before it was struck). In both sorts of cases, we say that the explanatory arguments in question are *defeated*.

**Definition 1** (Defeasibility).  $\Gamma \vdash \Delta$  is *defeasible* if and only if, for some formula  $A$ ,  $\Gamma, A \not\vdash \Delta$ .

Second, explanatory arguments are *minimal* in the sense that all of their premises are needed to infer their conclusion. More precisely, minimality holds that no logical consequence of the premises of an explanatory argument is explanatory unless it is logically equivalent to those premises.

**Definition 2** (Minimality).  $\Gamma \vdash \Delta$  is *minimal* if and only if (if  $\Gamma' \vdash \Delta$  and  $\Gamma \vdash \Gamma'$ , then  $\Gamma' \vdash \Gamma$ ).

In general, minimality prohibits the addition of irrelevant information to explanatory arguments. For example, if the liquid’s acidity explains why the litmus paper turns red, then it does not follow that the liquid’s acidity *and its potability* explains why the litmus paper turned red, even if the liquid is in fact potable.

Third, since contradictions do not explain, explanatory arguments must have consistent premises. Naturally, this means that the consequence relation underlying explanatory arguments must not validate the principle of *ex contradictione quodlibet* (ECQ)—i.e.  $A, \neg A \vdash B$  for any arbitrary  $B$ . There are numerous logical systems that renounce ECQ—most notably paraconsistent systems. However, the latter are typically designed to represent reasoning that is *tolerant* of inconsistency and that may, in some cases, even validate contradictions. By contrast, we must aim, in the opposite direction, to represent reasoning that is *intolerant* of inconsistency and contradiction. Let us call a logic that invalidates any reasoning that proceeds from inconsistent or contradictory premises *premise-consistent*. Since contradictions never explain, explanatory arguments are premise-consistent, not paraconsistent.

**Definition 3** (Premise-consistency).  $\Gamma \vdash \Delta$  is *premise-consistent* if and only if  $\Gamma \not\vdash$ .

Fourth, explanatory arguments are *irreflexive*. For instance, “the litmus paper turned red because it turned red” is not an explanation. More generally, no explanatory arguments should be of the form  $A \vdash A$ . Even partial self-explanations seem unacceptable;  $A \wedge B \vdash A$  and its ilk should also be prohibited from qualifying as explanatory arguments. Like ECQ, reflexivity partly defines the consequence relation of classical logic. However, while many logics abandon ECQ, few logics abandon the principle of reflexivity, let alone try to preserve the property of irreflexivity.<sup>4</sup> One particular difficulty arises from the fact that irreflexive logics cannot easily recover Modus Ponens (MP), yet many explanations require MP. Our irreflexive system preserves MP as a theorem.

**Definition 4.** Let  $\hat{\Gamma}$  denote the smallest set containing  $\Gamma$  that has the following property: If  $A \wedge B \in \hat{\Gamma}$ , then  $A, B \in \hat{\Gamma}$ .

**Definition 5** (Irreflexivity).  $\Gamma \vdash \Delta$  is *irreflexive* if and only if  $\hat{\Gamma} \cap \Delta = \emptyset$ .

As will become clear, our characterization of explanatory arguments requires us to compare defeasible inferences that are premise-consistent and irreflexive. In order to facilitate this comparison, we introduce the following definition.

**Definition 6** (Non-Triviality). An inference  $\Gamma \sim \Delta$  is *non-trivial* if and only if it is defeasible, premise-consistent, and irreflexive.

Fifth, explanations are *stable*, which philosophers of science have analyzed in different ways (Hempel 1965; Lange 2009; Mitchell 2003; Skyrms 1980; Woodward 2003). In its most general form,  $X$  is said to be stable if  $X$  remains unchanged as other conditions  $C$  change. For instance, suppose that a patient’s rash is explained by a particular bacterial infection, though an alternative potential explanation is that the rash is caused by an allergic reaction. The explanation is stable insofar as she would have a rash regardless of whether she had had an allergic reaction. Typically, the fundamental bearers of stability are taken to be laws or generalizations. By contrast, we take explanatory arguments to be the fundamental bearers of stability. Elsewhere we have developed a distinctively brand of stability that we call “sturdiness.” For the purposes of this paper, we only require a sketch of this concept.

Sturdiness is a comparative property among non-trivial inferences. A non-trivial inference is sturdy just in case it succeeds when all other non-trivial inferences that share its conclusion fail, where by ‘failure’ we mean *defeat* and by ‘success’, the absence of failure. More precisely, the failure we have in mind is that which obtains when the premises of an inference are false. Since premise-consistent inferences do not hold when their premises are false, a sturdy inference is one that remains undefeated when the premises of its competitors—i.e. those non-trivial inferences that share its conclusion—are false. It should be clear from what has been said thus far that representing sturdiness requires us to treat premise-*inconsistency* as a form of defeat. As we shall see, our formalism enables us to do just this.

Sixth, explanations are *detachable*. If the bacterial infection is the best explanation of the rash, then we may conclude that the patient has a bacterial infection. By detaching the explanans, we may make further predictions, design interventions, and construct new models. This is the animating idea behind IBE.

<sup>JM</sup>It is often held that for lower-grade patterns of abduction, such as IPE, the conclusion has a provisional status. But in the case of IBE, it seems that

<sup>JM</sup>Transition

## 4 The Logic of Explanatory Arguments and IBE

It what follows, we develop sequent calculi for explanatory arguments, IBE, and the BEW-operator. We begin by introducing our base logic LEA, defined over the standard propositional language,  $\mathcal{L}$ . LEA is composed of two parts. The first part,  $LK^\ominus$ , is based on a variant of Gentzen’s sequent calculus for classical logic that was proposed by Piazza and Pulcini (2015) with the aim of representing nonmonotonicity in terms of an inference’s context-sensitivity. Although we have altered most of the definitions, terminology, and rules with which Piazza and Pulcini (2015) introduced their calculus, we retain the key technical features of  $LK^S$  and inherit their proof of the cut-elimination theorem. Our modifications are intended, on one

<sup>4</sup>While there have been some logics in which reflexivity fails, irreflexivity holds in very few. Notable exceptions include input-output logics, logics of grounding, and quantum logics.

<sup>0</sup>Logicians of inferentialist and proof-theoretic persuasion have already explored systems in which transitivity fails (Ripley 2011; Tennant 2014; Hlobil 2016),

hand, to represent a concept of *defeat* more appropriate to the behavior of explanations, and, on the other, to extend the axioms of the system to non-logical, material inferences. The second part of the base system,  $LE^\blacktriangleright$  introduces an additional class of sequents and rules that govern their interaction with those of  $LK^\ominus$ , including a rule for abduction or IBE. Finally, we describe a system  $LEA^+$  defined over  $\mathcal{L}_\blacktriangleright$  that extends  $\mathcal{L}$  to include an object-language connective,  $\blacktriangleright$ , that makes commitments to explanatory arguments explicit.

The need for a base logic that includes two consequence relations stems from the following observation: If the BEW-operator encodes explanatory arguments, and if IBE is a legitimate rule of inference, then IBE cannot itself be an explanatory argument. Here is the support for this claim. Suppose that “*A best explains why B*” says that there is an explanatory argument from *A* to *B*. Now suppose (for *reductio*) that IBE is also an explanatory argument. From these suppositions it follows that the *best-explains-why*-connective expresses an explanatory argument from *B* to *A*, since this is the only way for there to be a good explanatory argument with the form of IBE: *A best explains why B*,  $B / \therefore A$ . But it also follows that such an inference can itself be made explicit by the *best-explains-why*-connective, thereby yielding the sentence  $[(A \text{ best explains why } B) \wedge B] \text{ best explains why } A$ . In so far as sentences of this form are even intelligible, it is far from obvious that they *claim* what an application of IBE *shows*. Thus, we should not conclude that IBE is an explanatory argument. Rather, IBE is one sort of inference and that made explicit by explanatory vocabulary is another.<sup>5</sup> This means that any logical system designed to represent both explanatory arguments and IBE must appeal to two distinct consequence relations.

In recognition of this point, LEA contains two consequence relations, or, more precisely, two *classes* of consequence relations, for as we shall see the system contains infinitely many consequence relations. The class whose rules are given by  $LK^\ominus$  is intended to represent those sets of inferences to which both IBE and *candidate* explanatory arguments belong. The class of consequence relations introduced by  $LE^\blacktriangleright$  and denoted by  $\vdash^\blacktriangleright$  is supposed to capture the behavior of explanatory arguments themselves. The chief results of this section are (1) that the theorems of LEA constructed with  $\vdash^\blacktriangleright$  exhibit all of the properties associated with explanatory arguments (Theorem 2), (2) that LEA can be extended ( $LEA^+$ ) to include an object-language expression for making explicit commitments to explanatory arguments (Theorem 3), and (3) that this extension is conservative (Corollary 4.1).

In what follows, we assume a propositional language,  $\mathcal{L}$ , for classical logic, that consists of a countably infinite set of atomic sentences  $At = \{p_1, \dots, p_n, q_1, \dots, q_n\}$ , the binary connectives  $\wedge$  and  $\vee$ , and the unary connective  $\neg$ . Let  $A, B, C, D$  range over formulas; let  $\Gamma, \Delta, \Sigma, \Theta$  range over sets of formulas; let  $S, T, U$  range over sets of sets of formulas, and let  $X, Y, Z$  range over sets of atoms.

We begin by employing the standard sequent notations—e.g.  $\Gamma, A \vdash B, \Delta$ . Formulas on the left side of the turnstile are called the *antecedent*; on the right side they are called the *succedent*. Commas in the antecedent are read ‘conjunctively’ and those on the right are read ‘disjunctively’. The formula with the connective in a rule is the *principal* formula of that rule, and its components in the premises are the *active* formulas.

The sequents in our calculi depart from the standard form in two respects: just below our turnstile we add a set of formulas,  $\Theta$ , called a *defeater set* and to the far left of the turnstile we add another,  $\Sigma$ , called a *background set*.

$$\Sigma \mid \Gamma \vdash^\ominus \Delta$$

<sup>5</sup>Treatments of abduction in the computer science literature have long recognized this point, if only implicitly, insofar as they have viewed abduction as a form of (restricted) deduction in reverse.

Defeater sets contain information whose addition to the premises would defeat the inference represented by the sequent. Roughly put, a sequent is defeated whenever its antecedent or background set contains a formula that is logically equivalent to a subset of the defeater set. Thus, as the name suggests, defeater sets are sets of inference-defeaters.

As noted, we adopt and modify the sequents and rules of the calculus  $LK^S$  developed by Piazza and Pulcini (2015). The general idea captured by this calculus is that any application of the rules along a derivation ought to preserve not only the validity, but also the defeat-status of sequents. In our system, LEA, this means that for any derivation tree,  $\pi$ , constructed by recursive applications of the rules (excluding *cut*), the following holds: if a defeated sequent occurs in a branch of  $\pi$ , then all sequents below it are also defeated. By tracking the defeat of sequents, our proof theory can identify if and when a line of defeasible reasoning goes astray.

We interpret background sets as consisting of information that is available to a reasoner when she draws an inference, but which does not serve as a premise and from which the conclusion is not said to follow. Having such a device in our formalism enables us to capture an important aspect of defeasibility, namely, that the introduction of new information may jeopardize prior inferential commitments even when that information does not serve as fodder for new inferences in its own right. These sets also serve a technical role in our system since they often expand in the course of a derivation, picking up traces of those formulas shifted by the rules from the left to the right side of the turnstile. (See  $\neg \vdash$  in Figure 1). It is this latter feature which permits the implementation of a Gentzen-style normalization procedure and proof of cut-elimination. Indeed, by keeping a *record*, so to speak, of formulas that have moved from the antecedent in a premise to the succedent of a conclusion, background sets ensure that all the sequents in a cut-free derivation are undefeated just in case its end-sequent is undefeated.<sup>6</sup>

Because the provability of any sequent in LEA depends, in part, upon the contents of its defeater set, there is not one or two but  $\mathcal{P}(\mathcal{L})$ -many consequence relations represented by the calculi. Some are classical, i.e.  $\Theta = \emptyset$ ; many are non-monotonic, i.e.  $\Theta \neq \emptyset$ ; others defy even the most ubiquitous structural properties, such as reflexivity, i.e.  $\Delta \subseteq \Theta$ . By specifying the contents of defeater sets, the rules of our calculi are able to exploit this panoply so as to home in on the class of consequence relations that bears precisely those properties we associate with explanatory arguments.

Before delving into the details of the system, we offer an informal gloss on our defeasible sequents. As noted, we follow Piazza and Pulcini (2015) to the extent that proofs in LEA preserve both validity and undefeatedness. However, since the rules in our calculi, and those of  $LE^\blacktriangleright$  in particular, are not deductive, it cannot be deductive ‘validity’ that is preserved. Instead, we follow Brandom (2008) and treat the sequents that belong to a proof as inferences that preserve *entitlement*. In keeping with this view, we offer the following reading of the defeasible sequent above: *Anyone entitled to (every member of)  $\Gamma$  is entitled to (at least one member of)  $\Delta$ , given the background of  $\Sigma$ .*

<sup>JM</sup>Issues with normative pragmatics interpretations of multiple succedent sequents.

<sup>JM</sup>transition

**Definition 7** (Defeater sets, Background sets, Defeasible Sequents). Defeater sets are sets of formulas that defeat an inference (see below). Background sets are sets of formulas that represent the background knowledge of the inference. A defeasible sequent is a standard

<sup>6</sup>In Piazza and Pulcini (2015) what we call background sets are simply referred to as *repositories* and while they play the same technical role, they are not provided with an substantive interpretation.

sequent with a *background set*,  $\Sigma$ , and a *defeater set*,  $\Theta$ , attached:<sup>7</sup>

$$\Sigma \mid \Gamma \mid_{\Theta} \Delta$$

When no background sets have been specified (i.e.  $\Sigma = \emptyset$ ) we write:

$$\cdot \mid \Gamma \mid_{\Theta} \Delta$$

In order to state the conditions under which a defeasible sequent is defeated, we must introduce some preliminary concepts.

**Definition 8** ( $\mathcal{D}$ -Rules). Let  $\mathcal{D}$  be the following set of rules:

$$\mathcal{D}_{\wedge}: A \in \mathcal{L} \Rightarrow A \wedge B \in \mathcal{L}$$

$$\mathcal{D}_{\vee}: A \in \mathcal{L} \text{ and } B \in \mathcal{L} \Leftrightarrow A \vee B \in \mathcal{L}$$

$$\mathcal{D}_{\neg}: A \in \mathcal{L} \Leftrightarrow \neg\neg A \in \mathcal{L}$$

$$\mathcal{D}_{\neg\vee\wedge}: \neg A \vee \neg B \in \mathcal{L} \Leftrightarrow \neg(A \wedge B) \in \mathcal{L}$$

$$\mathcal{D}_{\neg\wedge\vee}: \neg A \wedge \neg B \in \mathcal{L} \Leftrightarrow \neg(A \vee B) \in \mathcal{L}$$

These rules ought to be rather familiar.  $\mathcal{D}_{\neg}$  is the principle of Double Negation and  $\mathcal{D}_{\neg\vee\wedge}/\mathcal{D}_{\neg\wedge\vee}$  are *de Morgan's Laws*. Read left to right,  $\mathcal{D}_{\vee}$  is Conjunction Introduction with  $\wedge$  substituted for  $\vee$  and similarly,  $\mathcal{D}_{\wedge}$  is just Disjunction Introduction with the converse substitution. The motivation for these substitutions as well as the peculiar lack of a biconditional in  $\mathcal{D}_{\wedge}$  will become clear in a moment.

**Definition 9** ( $\mathcal{D}(\Theta)$ ). Let  $\mathcal{D}(\Theta)$  be the closure of  $\Theta$  under the  $\mathcal{D}$ -rules as well as the commutativity, associativity, and distributivity of conjunction and disjunction, respectively.

**Definition 10** (Compatibility). A set of formulas,  $\Gamma$ , is said to be compatible with a defeater set,  $\Theta$ , just in case the conjunction of the members of  $\Gamma$  is not included in  $\mathcal{D}(\Theta)$ . We use ' $\succsim$ ' to symbolize compatibility.

$$\Gamma \succsim \Theta \text{ iff } \bigwedge \Gamma \notin \mathcal{D}(\Theta)$$

**Example 1.**  $\{p_1 \vee q_1, \neg p_2, p_3 \wedge q_2\} \succsim \{p_1, p_2, \neg p_3 \vee \neg q_1\}$

**Example 2.**  $\{\neg(p_1 \vee q_1), \neg\neg\neg p_2, \neg p_3 \wedge q_2\} \succsim \{p_1 \wedge p_2, p_3, \neg q_2\}$

**Example 3.**  $\{\neg p_1 \vee q_1, \neg p_2, \neg p_3 \wedge q_2\} \not\succsim \{q_2\}$

**Example 4.**  $\{\neg(p_1 \vee q_1), \neg\neg\neg p_2, \neg p_3 \wedge q_2\} \not\succsim \{\neg p_1, \neg p_2\}$

**Remark 1.** For any set of formulas  $\Gamma$ ,  $\Gamma \succsim \emptyset$ .

<sup>7</sup>In general, the respective roles played by *control sets*, *compatibility*, and *soundness* in Piazza and Pulcini's  $\text{LK}^S$  are played by *defeater sets*, (our notion of) *compatibility*, and *undefeatedness* in our system. The crucial difference between the two approaches is that while in Piazza and Pulcini (2015), the occurrence of a disjunctive formula in the antecedent would render the sequent defeated (resp. unsound) if either of the disjuncts occurred in the sequent's defeater set (resp. control set), in our system, the sequent would only be defeated if both disjuncts were present in the defeater set. We believe that the latter property captures a more intuitive, less cautious conception of defeat. Unfortunately, this conception of defeat could only be purchased at the cost of attaching a *proviso* to  $\vee \vdash$  that restricts the rule's application to sequents whose active formulas are compatible with their respective defeater sets. On balance, we were willing to trade the elegance of the rules for the more accurate portrayal of defeat. In order to realize this conception in our definitions, we found it necessary to deploy substantially different operations and have re-named the resultant concepts so as to avoid confusion. With this said, Definition 12 and Lemma 1 are taken over from Piazza and Pulcini (2015) with very little modification.



**Definition 11** (Defeat). A defeasible sequent,  $\Sigma \mid \Gamma \mid_{\Theta} \Delta$ , is said to be *undefeated* whenever  $\Sigma \cup \Gamma \succsim \Theta$  and defeated otherwise.

The  $\mathcal{D}$ -Rules are designed to generate closed sets of formulas whose occurrence in the antecedent defeats the sequent to which the defeater set is attached. Since a formula of the form  $A \wedge B$  defeats an inference just in case one or more of its constituents belongs to the defeater set,  $\mathcal{D}_{\wedge}$  is constructed so that a set closed under it will contain all those conjunctions for which at least one member of the original set,  $\Theta$ , is a conjunct. Conversely, if the defeater set contains a conjunction but neither of its conjuncts, then we know that the two formulas *together* defeat the sequent but not whether either formula by itself would be sufficient for defeat. (The need for defeat to occur when both conjuncts appear on their own in the antecedent is handled by the conjunctive formulation of *compatibility*.) Thus, unlike the other  $\mathcal{D}$ -rules,  $\mathcal{D}_{\wedge}$  only permits the construction of more complex formulas—hence the absence of a biconditional in its formulation.

In contrast, a disjunctive formula defeats an inference only when both of the disjuncts occur in its defeater set. The  $\mathcal{D}_{\vee}$ -rule captures this intuition—read left-to-right—by constructing disjunctions when both disjuncts are present in the defeater set. Conversely, a disjunction in a defeater set means that the presence of either disjunct in the antecedent will defeat the sequent. Thus, the  $\mathcal{D}_{\vee}$ -rule is formulated—from right-to-left—so that a disjunction in the  $\mathcal{D}$ -closure of a defeater set will contain both disjuncts. The result of comparing the antecedent (and background set) with the closure of the defeater set under the  $\mathcal{D}$ -Rules is a definition of compatibility that underwrites an intuitive conception of defeat. We illustrate the nature of compatibility with the following lemma.

**Lemma 1.** 1. If  $\Gamma \cup \Delta \succsim \Theta$  and  $\Lambda \subset \Theta$ , then  $\Delta \succsim \Lambda$ .

2.  $\Gamma \cup \{A \wedge B\} \succsim \Theta$  iff  $\Gamma \cup \{A, B\} \succsim \Theta$
3.  $\Gamma \cup \{A \vee B\} \succsim \Theta$  iff  $\Gamma \cup \{A\} \succsim \Theta$  or  $\Gamma \cup \{B\} \succsim \Theta$
4.  $\Gamma \cup \{\neg \neg A\} \succsim \Theta$  iff  $\Gamma \cup \{A\} \succsim \Theta$

*Proof.* For the first sub-theorem, suppose for *reductio* that  $\Delta \not\succsim \Lambda$ . It would then follow that  $\bigwedge \Delta \in \mathcal{D}(\Lambda)$ . But then, by hypothesis,  $\bigwedge \Delta \in \mathcal{D}(\Theta)$  and thus, against our assumption,  $\Gamma \cup \Delta \not\succsim \Theta$ . The remaining sub-theorems follow directly from Definitions 8, 9, and 10. ■

#### 4.1 $\text{LK}^{\Theta}$

Our notions of compatibility and defeat enable us to capture premise-consistency and defeasibility. To demonstrate this, we invite the reader to consider the rules for  $\text{LK}^{\Theta}$  in Figure 1. These rules are designed to generate trees that preserve validity *downward* and undefeat-*edness upward*. The most natural way to read the rules is bottom-up as follows: “If [the conclusion sequent] is undefeated, then so is/are [the premise sequent/s].” Alternatively, the rules may be read top-down as permitting the conclusion, given the premises, so long as the former is not defeated. On either reading, the rules are formulated to ensure that a cut-free derivation whose end-sequent is undefeated will contain only undefeated sequents throughout. This property is important both for the establishment of cut-eliminability and, more primitively, for the fact that proofs in a nonmonotonic system should not contain defeated sequents.

**Definition 12** (Proof, Paraproof). For a rooted, finitely branching tree  $\pi$  whose nodes are sequents of  $\text{LK}^{\Theta}$  (LEA), and which is recursively built up from axioms by means of the rules

of  $\text{LK}^\ominus$  (LEA), if each sequent in  $\pi$  is undefeated, then  $\pi$  is said to be a proof of  $\text{LK}^\ominus$  (LEA), otherwise  $\pi$  is called a paraproof.

The axioms of  $\text{LK}^\ominus$  come in two varieties. *Logical axioms* are the familiar ‘initial sequents’ of LK to which defeater sets are attached. *Proper axioms*, on the other hand, are sequents composed of nonempty, non-overlapping sets of atoms on the left and right of the turnstile.<sup>8</sup> These axioms are intended to represent the non-logical, material inferences that figure in various types of scientific reasoning. Since such inferences are often the products of concrete empirical inquiries, we insist that they be introduced with non-empty *background sets* of (possibly complex) formulas that reflect the epistemic context of their use.

In order to avoid having defeated axioms—i.e. non-starters—we must place certain constraints on the defeater sets of initial sequents.

**Definition 13** (Constraints on Defeater Sets for Axioms). If  $\cdot \mid p \mid_{\Theta} p$  is a logical axiom and  $\Sigma \mid X \mid_{\Psi} Y$  is a proper axiom, then

- (i)  $\Theta$  and  $\Psi$  are sets of literals.
- (ii)  $p \notin \Theta$  and  $\forall p \in X (p \notin \Psi)$
- (iii)  $\Sigma \cap \Psi = \emptyset$

The first constraint restricts the defeater sets of axioms to a low level of formula complexity in order to limit the lacuna that occurs when a conjunction but neither of its conjuncts belongs to a defeater set. The second ensures that initial sequents are not defeated by their antecedents and, in the case of logical axioms, that equivalence among atoms is preserved. While this move protects reflexivity in  $\text{LK}^\ominus$ , the rules of  $\text{LE}^\blacktriangleright$  will prevent this property from being transferred to properly explanatory arguments. Finally, the third constraint prevents proper axioms from being defeated by their initial background sets.

Before demonstrating that these constraints allow us to produce a nonmonotonic system with premise-consistent theorems, we pause to explain some of the more exotic features of  $\text{LK}^\ominus$ . First, note that, with the exception of DE, single-premise rules have no effect on defeater sets, whereas, all the two-premise rules yield defeater sets in the conclusion that are the union of those in the premises. This fact guarantees that no information about potential defeaters is lost along a derivation.

**Proposition 1. A.** If  $\pi$  is a tree whose root is  $\Sigma \mid \Gamma \mid_{\Theta'} \Delta$  and  $\cdot \mid p \mid_{\Theta} p$  occurs in the leaves of  $\pi$  and  $A \in \Theta$ , then  $A \in \Theta'$ .

**B.** If  $\pi$  is a tree whose root is  $\Sigma \mid \Gamma \mid_{\Theta'} \Delta$  and  $\Sigma \mid X \mid_{\Theta} Y$  occurs in the leaves of  $\pi$  and  $A \in \Theta$ , then  $A \in \Theta'$ .

*Proof.* Follows directly from the rules of  $\text{LK}^\ominus$ . ■

Exempting DE from this requirement is justified by the idea that reasoners ought to be able to add new, extra-logical information about defeaters to their arguments as it becomes available. The DE rule (which stands for *Defeater Expansion*) allows one to do so as long as it does not defeat the sequent in question.

<sup>8</sup>It is well-known that the addition of non-tautological axioms to the system of classical logic will lead to inconsistency if those axioms are taken to be closed under universal substitution (US). Even if closure under US is abandoned for proper axioms, their addition to a sequent system such as LK, threatens the cut-elimination theorem. Fortunately, Piazza and Pulcini (2016) have shown how to generate non-logical axiomatic extensions of classical propositional logic that admit cut elimination. While such extensions are obviously not complete—they are *post-competes*—axioms can be formulated so as to preserve consistency. The trick to doing so is to ensure that the empty sequent does belong to the set of proper axioms (See Theorem 3.7 in Piazza and Pulcini (2016)). Our proper axioms have been formulated in conformity with this constraint.

### Logical Axioms

$$\frac{}{\cdot \mid p \mid_{\Theta} p} \text{ log. ax.}$$

### Proper Axioms

$$\frac{}{\Sigma \mid X \mid_{\Theta} Y} \quad \Sigma, X, Y \text{ are nonempty; } X \cap Y = \emptyset \quad \text{prop. ax.}$$

### Cut Rule

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} A, \Delta \quad \Sigma' \mid \Gamma', A \mid_{\Psi} \Delta'}{\Sigma', \Sigma \mid \Gamma', \Gamma \mid_{\Theta \cup \Psi} \Delta, \Delta'} \text{ cut}$$

### Structural Rules

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} \Delta}{\Sigma \mid \Gamma, A \mid_{\Theta} \Delta} \text{ LW}$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} \Delta}{\Sigma \mid \Gamma \mid_{\Theta} \Delta, A} \text{ RW}$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} \Delta}{\Sigma \mid \Gamma \mid_{\Theta \cup \Psi} \Delta} \text{ DE}$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} \Delta}{\Sigma, A \mid \Gamma \mid_{\Theta} \Delta} \text{ BE}$$

### Logical Rules

$$\frac{\Sigma \mid \Gamma, A, B \mid_{\Theta} \Delta}{\Sigma \mid \Gamma, A \wedge B \mid_{\Theta} \Delta} \wedge \vdash$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} A, \Delta \quad \Sigma' \mid \Gamma' \mid_{\Psi} B, \Delta'}{\Sigma', \Sigma \mid \Gamma', \Gamma \mid_{\Theta \cup \Psi} A \wedge B, \Delta, \Delta'} \vdash \wedge$$

$$\frac{\Sigma \mid \Gamma, A \mid_{\Theta} \Delta \quad \Sigma' \mid \Gamma', B \mid_{\Psi} \Delta'}{\Sigma', \Sigma \mid \Gamma', \Gamma, A \vee B \mid_{\Theta \cup \Psi} \Delta, \Delta'} \vee \vdash^{\dagger}$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} A, B, \Delta}{\Sigma \mid \Gamma \mid_{\Theta} A \vee B, \Delta} \vdash \vee$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta} A, \Delta}{\Sigma \mid \Gamma, \neg A \mid_{\Theta} \Delta} \neg \vdash$$

$$\frac{\Sigma \mid \Gamma, A \mid_{\Theta} \Delta}{\Sigma, A \mid \Gamma \mid_{\Theta} \neg A, \Delta} \vdash \neg$$

$\dagger$  Provided that  $\{A\} \lesssim \Theta$  and  $\{B\} \lesssim \Psi$ .

**Figure 1:** Rules for  $\text{LK}^{\Theta}$

Second, all of the rules either transfer or combine background sets from premises to conclusions, except for BE and  $\vdash \neg$ . The former (whose name abbreviates *Background Expansion*) permits the addition of arbitrary formulas to a sequent's background set. At one level, this rule simply captures the way new contextual information is added in the course of scientific reasoning. But at a deeper level, it attempts to represent the way reasoners might *probe* the defeasibility of an inference by discharging it in different contexts. In this sense, BE codifies certain patterns of *experimental* reasoning.

On the other hand  $\vdash \neg$  adds the (active) antecedent of the premise to the background set of the conclusion. This behavior is of a piece with the explanation given for background sets above—they act as a kind of *record* of those formulas that have been shifted from the left to the right of a turnstile.<sup>9</sup> An informal interpretation of the rule (read upwards) can be given as follows: if one is entitled to  $\neg A$  while  $A$  is in one's background set of entitlements, then one is entitled to whatever follows from  $A$  once it has been removed from that background set and entitlement to its negation has been renounced. The formulation of  $\vdash \neg$  in this manner is critical to the upwards preservation of undefeatedness in cut-free proofs.

**Proposition 2.** Any cut-free paraproof in  $LK^\ominus$  is a proof if and only if its end-sequent is undefeated, i.e. undefeatedness is preserved upwards in cut-free proofs.

*Proof.* For all of the rules except  $\vee \vdash$ , the undefeatedness of the premises follows directly from that of the conclusion by way of Lemma 1.1, 1.2 or 1.4. In the case of  $\vee \vdash$ , it follows from Lemma 1.3 that it would be possible for the conclusion to be undefeated while exactly one of the premises is defeated. This possibility, however, is blocked by the proviso ( $\dagger$ ) that restricts the rule's application to sequents whose active formulas are compatible with their respective defeater sets. ■

The preservation of undefeatedness in cut-free proofs is, in turn, critical to cut-elimination because by permuting cut upwards in derivations, the normalization procedure occasionally turns proofs into paraproofs.<sup>10</sup> Proposition 2, however, ensures that for any such paraproof, there is a cut-free proof of its end-sequent.

**Lemma 2.** Any sequent that is provable in  $LK^\ominus$  has a cut-free proof.

*Proof.* As noted above, the similarity between our  $LK^\ominus$  and Piazza et al.'s  $LK^S$  enables us to inherit the latter's cut-elimination theorem. We refrain from presenting the proof of the theorem here, but the interested reader is invited to see Piazza and Pulcini (2015) Theorem 19. ■

## 4.2 $LE^\blacktriangleright$ and LEA

With the rules for  $LK^\ominus$  now in place, we can turn to the other part of our base system,  $LE^\blacktriangleright$ , where we introduce a class of consequence relations, denoted by  $\vdash_\ominus^\blacktriangleright$ , that represents explanatory arguments. We will call sequents constructed with this turnstile  $\blacktriangleright$ -sequents. Before discussing the rules for this system, we need a few preliminary concepts.

**Definition 14** ( $\neg(\Gamma)$ ). Let  $\neg(\Gamma)$  stand for the following operation:

$$\neg(\Gamma) =_{df} \{ \neg A : A \in \Gamma \}$$

<sup>9</sup>By this same reasoning, the *cut* rule also ought to add the active formula in its premises to the background set of the conclusion. However, the *cut* rule is exempted on the grounds that such a rule is just a statement about the conditions under which information may be *removed* from a proof.

<sup>10</sup>See Piazza and Pulcini (2015, 19) for an example.

The operation easily extends to sets of sets:

$$\neg(\mathbf{S}) =_{df} \{\neg(\Gamma) : \Gamma \in \mathbf{S}\}$$

**Definition 15** ( $\bar{\Gamma}$ ). Let  $\bar{\Gamma}$  stand for the following set function:

$$\bar{\Gamma} =_{df} \neg(\Gamma) \cup \{A \wedge \neg A : A \in \mathcal{L}\}$$

By definition,  $\mathcal{D}(\bar{\Gamma})$  contains all of the contradictory formulas in  $\mathcal{L}$ . This set function will thus be instrumental to securing premise-consistency for  $\blacktriangleright$ -sequents.

**Definition 16** (Competitor Set,  $\mathbf{S}^\Delta$ ). Let  $\mathbf{S}^\Delta$  stand for the following set:

$$\mathbf{S}^\Delta =_{df} \{\Omega_i : \Sigma_i \mid \Omega_i \mid_{\Psi_i \cup \bar{\Omega}_i \cup \Delta}^{\text{LK}^\Theta} \Delta\}$$

where  $i \in \mathbb{N}$  and  $\text{LK}^\Theta$  above a turnstile indicates that the sequent is a theorem of  $\text{LK}^\Theta$ .

Despite its complex appearance,  $\mathbf{S}^\Delta$  is nothing more than the set containing the antecedents of all those nontrivial (i.e. premise-consistent and irreflexive) theorems of  $\text{LK}^\Theta$  that have  $\Delta$  as its conclusion. For ease of reference, the definition indexes background sets and defeater sets by the antecedents of the relevant sequent. This definition is crucial to securing sturdiness.

**Definition 17** (LK-Equivalence Set,  $\mathbf{T}_\Gamma$ ). Let  $\mathbf{T}_\Gamma$  stand for the following set:

$$\mathbf{T}_\Gamma =_{df} \{\Lambda : \Gamma \mid_{\text{LK}} \Lambda \text{ and } \Lambda \mid_{\text{LK}} \Gamma\}$$

where  $\text{LK}$  above a turnstile indicates that the sequent is a theorem of  $\text{LK}$ .

The members of  $\mathbf{T}_\Gamma$  are just those sets that are provably equivalent to  $\Gamma$  in classical logic, i.e.  $\text{LK}$ .

**Definition 18** (Disjunction Deletion,  $\lfloor \mathbf{S} \rfloor$ ).

$$\lfloor \mathbf{S} \rfloor =_{df} \{\Delta \setminus \{A \vee B\} : \Delta \in \mathbf{S} \text{ and } A \in \bigcup \mathbf{S}\}$$

The set denoted by  $\lfloor \mathbf{S} \rfloor$  is the result of deleting from the members of  $\mathbf{S}$  any disjunction one of whose disjuncts belongs to a member of  $\mathbf{S}$ .

**Definition 19** ( $\Delta \parallel \Gamma$ ).

$$\Delta \parallel \Gamma =_{df} \begin{cases} \Delta & \text{if } \Delta \subseteq \Gamma \\ \Delta \setminus \Gamma & \text{otherwise} \end{cases}$$

We extend the definition to sets of sets as follows:

$$\mathbf{S} \parallel \mathbf{T} =_{df} \{\Delta \parallel \Gamma : \Delta \in \mathbf{S} \text{ and } \Gamma \in \mathbf{T}\}$$

The set denoted by  $\mathbf{S} \parallel \mathbf{T}_\Gamma$  consists of those members of  $\mathbf{S}$  that are subsets of  $\Gamma$  (and its classically provable equivalents) and the set-complements relative to  $\Gamma$  (and its classically provable equivalents) of those members that are not subsets of  $\Gamma$  (or its classically provable equivalents). As we shall see, Definitions 18 and 19 are needed to secure minimality.

We will now consider the rules for  $\text{LE}^\blacktriangleright$ , presented in Figure 2. Since the structural rules for  $\text{LE}^\blacktriangleright$  are just the **DE** and **BE** rules, *modulo* the  $\blacktriangleright$  turnstile, we begin with what we have called the *Mixed Rules*. These are both the most unusual—as they contain two different types

### Structural Rules

$$\frac{\Sigma \mid \Gamma \mid_{\Theta}^{\blacktriangleright} \Delta}{\Sigma \mid \Gamma \mid_{\Theta \cup \Psi}^{\blacktriangleright} \Delta} \text{DE}^{\blacktriangleright}$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta}^{\blacktriangleright} \Delta}{\Sigma, A \mid \Gamma \mid_{\Theta}^{\blacktriangleright} \Delta} \text{BE}^{\blacktriangleright}$$

### Mixed Rules

$$\frac{\bigcup \neg([S^{\Delta}] \parallel T_{\Gamma}), \Sigma \mid \Gamma \mid_{\Theta \cup \bar{\Gamma} \cup \Delta}^{\blacktriangleright} \Delta}{\Sigma \mid \Gamma \mid_{\Theta \cup \bar{\Gamma} \cup \Delta}^{\blacktriangleright} \Delta} \text{sturdy}^{\ddagger}$$

$$\frac{\Sigma \mid \Gamma, A \mid_{\Theta}^{\blacktriangleright} \Delta \quad \Sigma' \mid \Gamma' \mid_{\Psi \cup \bar{\Gamma}' \cup \Delta}^{\blacktriangleright} \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \mid_{\Theta \cup \Psi \cup \bar{\Gamma}' \cup \Delta}^{\blacktriangleright} A} \text{abduct.}$$

### Logical Rules

$$\frac{\Sigma \mid \Gamma, A, B \mid_{\Theta}^{\blacktriangleright} \Delta}{\Sigma \mid \Gamma, A \wedge B \mid_{\Theta}^{\blacktriangleright} \Delta} \wedge \vdash^{\blacktriangleright}$$

$$\frac{\Sigma \mid \Gamma \mid_{\Theta}^{\blacktriangleright} A, \Delta \quad \Sigma' \mid \Gamma \mid_{\Psi}^{\blacktriangleright} B, \Delta}{\Sigma', \Sigma \mid \Gamma \mid_{\Theta \cup \Psi}^{\blacktriangleright} A \wedge B, \Delta} \vdash^{\blacktriangleright} \wedge$$

$\ddagger$  Provided that (i) all sets are nonempty, (ii)  $\Sigma \subseteq \Sigma_i$ , (iii)  $\Gamma \cap \Sigma_i = \emptyset$ , (iv)  $\Omega_i \cap \Sigma = \emptyset$ .

$\star$  Provided that  $\Gamma \cap \Gamma' \cap \{A\} = \emptyset$ .

**Figure 2:** Rules for  $\text{LE}^{\blacktriangleright}$

consequence relations—and the most important for the purposes of representing explanatory arguments.

As mentioned above, *sturdiness* is our proposal for how to understand the property of *stability* associated with explanations. In slogan form, inferences are sturdy just in case they succeed where all others fail. To say that one inference succeeds when another fails, we can imagine the following procedure:

**Step 1:** Line up all of the nontrivial inferences that have the explanandum,  $B$ , as their conclusion. For each of these inferences, all other nontrivial inferences leading to  $B$  are its “competitors.”

**Step 2:** For each  $A$  that has  $B$  as a nontrivial consequence, suppose that all of  $A$ ’s competitors’ premises are false.

**Step 3:** If the falsehood of any of these competitors defeats the inference from  $A$  to  $B$ , then the latter is not sturdy; otherwise, it is sturdy.

Our *sturdy* rule aims to formalize this procedure. The first step is represented by the fact that the succedent of the premise is  $\Delta$  and its background set includes a set obtained from  $S^{\Delta}$ . The second step is captured by the fact that  $\bigcup \neg([S^{\Delta}] \parallel T_{\Gamma})$  appears in the background set of the premise. Roughly put, this set contains the negations of all the antecedents of nontrivial inferences whose succedent is  $\Delta$ , with the caveat that the members of  $\Gamma$  and their equivalents are removed from any antecedent that is a superset of  $\Gamma$ . Finally, the third step is reached when the premise sequent carries down into the conclusion where it appears with the explanatory-argument-denoting turnstile, i.e.  $\mid_{\Theta}^{\blacktriangleright}$ .

By restricting premise sequents to those whose defeater sets contain  $\bar{\Gamma}$ —where  $\Gamma$  is the antecedent—the *sturdy* rule ensures that the antecedents of  $\blacktriangleright$ -sequents are consistent and

thereby enables these sequents to represent premise-consistent inferences.

**Lemma 3.** The sequent  $\Sigma \mid \Gamma, A, \neg A \frac{\triangleright}{\Theta} \Delta$  is not a theorem of LEA.

*Proof.* Suppose for *reductio* that  $\Sigma \mid \Gamma, A, \neg A \frac{\triangleright}{\Theta} \Delta$  is a theorem and hence is undefeated. It must be derived via an application of *sturdy* whose premise, if we omit the background set, is  $\Gamma, A, \neg A \frac{}{\Theta \cup \{A, \neg A\} \cup \Delta} \Delta$ . By Definition 15,  $\bigwedge \{A, \neg A\} \in \mathcal{D}(\{A, \neg A\})$ , since  $A \wedge \neg A \in \{B \wedge \neg B : B \in \mathcal{L}\}$ . Thus, *contra* our supposition,  $\Sigma \mid \Gamma, A, \neg A \frac{\triangleright}{\Theta} \Delta$  is defeated. ■

**Lemma 4.** The sequent  $\Sigma \mid \Gamma, A \wedge \neg A \frac{\triangleright}{\Theta} \Delta$  is not a theorem of LEA.

*Proof.* Same as for Lemma 3. ■

The *sturdy* rule also prevents  $\triangleright$ -sequents from being reflexive. It does so by restricting the defeater sets of premise sequents to those which contain the succedent. We can now show that both partial and complete self-explanations will be prohibited from  $\text{LE}^\triangleright$ .

**Lemma 5.** The sequent  $\Sigma \mid \Gamma, A \frac{\triangleright}{\Theta} A, \Delta$  is not a theorem of LEA.

*Proof.* Suppose for *reductio* that  $\Sigma \mid \Gamma, A \frac{\triangleright}{\Theta} A, \Delta$  is a theorem and hence is undefeated. It must be derived via an application of *sturdy* whose premise, if we omit the background set, is  $\Gamma, A \frac{}{\Theta \cup \{A\} \cup \Delta} A, \Delta$ . But, since (by Definition 9)  $\bigwedge(\Gamma \cup \{A\}) \in \mathcal{D}(A)$ , this sequent is defeated, contradicting our supposition. ■

**Lemma 6.** The sequent  $\Sigma \mid \Gamma, A \wedge B \frac{\triangleright}{\Theta} A, \Delta$  is not a theorem of LEA.

*Proof.* Same as for Lemma 5. ■

Furthermore, neither premise-consistency, nor irreflexivity prevent instances of MP from standing as candidates for explanatory arguments, i.e. as premises of *sturdy*. For the remaining proofs, we omit arbitrary background sets (i.e.  $\Sigma$ ) whenever possible.

**Proposition 3.** If the sequent  $\Sigma \mid A, \neg A \vee B \frac{}{\Theta} B$  is a theorem of LEA, then so is  $\Sigma \mid A, \neg A \vee B \frac{}{\Theta \cup \{A, \neg A \vee B\} \cup \{B\}} B$ .

*Proof.* Suppose for *reductio* that there is no proof  $\pi$  of  $A, \neg A \vee B \frac{}{\Theta \cup \{A, \neg A \vee B\} \cup \{B\}} B$  from  $A, \neg A \vee B \frac{}{\Theta} B$ . It follows that either  $\pi$  is not a paraproof or  $A, \neg A \vee B \frac{}{\Theta \cup \{A, \neg A \vee B\} \cup \{B\}} B$  is defeated. But  $A, \neg A \vee B \frac{}{\Theta \cup \{A, \neg A \vee B\} \cup \{B\}} B$  follows from  $A, \neg A \vee B \frac{}{\Theta} B$  by a single application of DE, and thus  $\pi$  is at least a paraproof. If  $A, \neg A \vee B \frac{}{\Theta \cup \{A, \neg A \vee B\} \cup \{B\}} B$  is defeated, then  $\bigwedge \{A, \neg A \vee B\} \in \mathcal{D}(\Theta \cup \{A, \neg A \vee B\} \cup \{B\})$ . But  $A \wedge (\neg A \vee B) \notin \mathcal{D}(\{B\})$ , and by Definition 9 and 15 it follows that  $A \wedge (\neg A \vee B) \notin \mathcal{D}(\{A, \neg A \vee B\})$ . Therefore, it must be the case that  $A \wedge (\neg A \vee B) \in \mathcal{D}(\Theta)$ . But by hypothesis  $A, \neg A \vee B \frac{}{\Theta} B$  is undefeated. Thus, *contra* our supposition, there is a proof of  $A, \neg A \vee B \frac{}{\Theta \cup \{A, \neg A \vee B\} \cup \{B\}} B$ . ■

Less obvious than the achievement of irreflexivity is the fact that *sturdy* also ensures that  $\triangleright$ -sequents are minimal in the sense articulated above by Definition 2. To prove this, we begin by demonstrating that no proper subset of the premises of an explanatory argument is itself explanatory of the same explanandum. This ensures that the condition in Definition 2 is trivially satisfied.

**Lemma 7.** If the sequent  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} \Delta$  is a theorem of LEA, then  $\Sigma \mid \Gamma \mid_{\Theta}^{\triangleright} \Delta$  is not.

*Proof.* We proceed by proving the contrapositive. Suppose for conditional proof that  $\Gamma \mid_{\Theta}^{\triangleright} \Delta$  is a theorem of LEA. It follows from *sturdy* that  $\Gamma \mid_{\Theta}^{\triangleright} \Delta$  is a theorem and from Definition 16 that  $\Gamma \in [\mathbf{S}^{\Delta}]$ . Next suppose for *reductio* that  $\Gamma, A \mid_{\Theta}^{\triangleright} \Delta$  is also a theorem. From *sturdy* it follows that  $\bigcup \neg([\mathbf{S}^{\Delta}] \parallel \mathbf{T}_{\Gamma \cup \{A\}}) \mid \Gamma, A \mid_{\Theta \cup \overline{\Gamma \cup \{A\}} \cup \Delta}^{\triangleright} \Delta$  is a theorem. Since  $\Gamma \subseteq \Gamma \cup \{A\}$ , it follows from Definition 16 and 19 that  $\neg(\Gamma) \subset \bigcup \neg([\mathbf{S}^{\Delta}] \parallel \mathbf{T}_{\Gamma \cup \{A\}})$ . But from Definition 9 it follows that  $\bigwedge \neg(\Gamma) \in \mathcal{D}(\overline{\Gamma \cup \{A\}})$ , so, according to Definition 10  $\bigcup \neg([\mathbf{S}^{\Delta}] \parallel \mathbf{T}_{\Gamma \cup \{A\}}) \mid \Gamma, A \mid_{\Theta \cup \overline{\Gamma \cup \{A\}} \cup \Delta}^{\triangleright} \Delta$  is defeated, contradicting our supposition. Thus, if  $\Gamma \mid_{\Theta}^{\triangleright} \Delta$  is a theorem, then  $\Gamma, A \mid_{\Theta}^{\triangleright} \Delta$  is not. It follows by contraposition and double negation that if the sequent  $\Gamma, A \mid_{\Theta}^{\triangleright} \Delta$  is a theorem of LEA, then  $\Gamma \mid_{\Theta}^{\triangleright} \Delta$  is not. ■

Next, we show that a disjunctive candidate is never sturdy when it must compete against one of its disjuncts. Again, this ensures that the condition in Definition 2 is trivially satisfied when  $\Gamma' = \Gamma \cup \{A \vee B\} \setminus \{A\}$ .

**Lemma 8.** If the sequent  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} \Delta$  is a theorem of LEA, then  $\Sigma \mid \Gamma, A \vee B \mid_{\Theta'}^{\triangleright} \Delta$  is not.

*Proof.* Suppose for *reductio* that  $\Gamma, A \vee B \mid_{\Theta'}^{\triangleright} \Delta$  is a theorem. The latter must be obtained by an application of *sturdy* preceded by an application of **BE** to  $\Gamma, A \vee B \mid_{\Theta''}^{\triangleright} \Delta$  (we omit the background set  $\Sigma$ ). If  $\pi$  is the branch immediately above  $\Gamma, A \vee B \mid_{\Theta''}^{\triangleright} \Delta$ , then it follows from  $\vee \vdash$  that  $\Gamma', A \mid_{\Theta'''}^{\triangleright} \Delta'$  and  $\Gamma'', B \mid_{\Theta'''}^{\triangleright} \Delta''$  must appear in  $\pi$ . By hypothesis and *sturdy* we know that  $\Gamma, A \mid_{\Theta}^{\triangleright} \Delta$ , and since  $\Delta = \Delta' \cup \Delta''$  from  $\vee \vdash$ , it must be the case that  $\Delta' = \Delta$  and  $\Delta'' = \emptyset$ . Moreover, if  $\Delta'' = \emptyset$ , then  $\Gamma'', B \mid_{\Theta'''}^{\triangleright} \emptyset$ , and thus there is some formula  $C \in \Gamma''$  such that  $C \vdash \neg B$ . Since  $\Gamma = \Gamma' \cup \Gamma''$  from  $\vee \vdash$ , we know that  $C \in \Gamma$ . According to our supposition,  $\bigcup \neg([\mathbf{S}^{\Delta}] \parallel \mathbf{T}_{\Gamma \cup \{A \vee B\}}), \Sigma \mid \Gamma, A \vee B \mid_{\Theta'}^{\triangleright} \Delta$  is a theorem and  $\Theta' = \Theta'' \cup \overline{\Gamma \cup \{A \vee B\}} \cup \Delta$ . We know that  $\{\neg A \wedge \neg B\} \in \overline{\Gamma \cup \{A \vee B\}}$ . Since  $C \in \Gamma$  and  $C \vdash \neg B$  it follows from Definition 9 that  $C \in \mathcal{D}(\overline{\Gamma \cup \{A \vee B\}})$ . Now, by hypothesis,  $\{\neg A\} \in \neg([\mathbf{S}^{\Delta}] \parallel \mathbf{T}_{\Gamma \cup \{A \vee B\}})$ . We can conclude that  $\neg A \wedge \neg C$  is in both  $\mathcal{D}(\Theta')$  as well as the conjunction of the background set and antecedent of  $\bigcup \neg([\mathbf{S}^{\Delta}] \parallel \mathbf{T}_{\Gamma \cup \{A \vee B\}}), \Sigma \mid \Gamma, A \vee B \mid_{\Theta'}^{\triangleright} \Delta$ , and, therefore, that they are incompatible. Since this sequent is thus defeated,  $\Gamma, A \vee B \mid_{\Theta'}^{\triangleright} \Delta$  is not a theorem, contradicting our supposition. ■

**Lemma 9.** If  $\Sigma \mid \Gamma, A, B \mid_{\Theta}^{\triangleright} \Delta$  is a theorem of LEA, then  $\Sigma \mid \Gamma, A \wedge B \mid_{\Theta'}^{\triangleright} \Delta$  is a theorem.

*Proof.* Follows Lemma 1.2, and *sturdy*. ■

**Lemma 10.** If  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} \Delta$  is a theorem of LEA, then  $\Sigma \mid \Gamma, \neg \neg A \mid_{\Theta'}^{\triangleright} \Delta$  is a theorem.

*Proof.* Follows Lemma 1.4, and *sturdy*. ■

**Lemma 11.** If  $\Sigma \mid \Gamma \mid_{\Theta}^{\triangleright} \Delta$  and  $\Sigma \mid \Gamma' \mid_{\Theta'}^{\triangleright} \Delta$  are theorems of LEA and  $\Gamma \vdash \Gamma'$ , then  $\Gamma' \vdash \Gamma$ .

*Proof.* There are five cases to consider. In the first case, where  $\Gamma' \subset \Gamma$ , Lemma 7 shows that the conditional is trivially satisfied. In the second case, where  $\Gamma' = \Gamma \cup \{A \wedge B\} \setminus \{A, B\}$  or  $\Gamma' = \Gamma \cup \{\neg \neg A\} \setminus \{A\}$ , the conditional follows from Lemmas 9 and 10, respectively. In the third case, where  $\Gamma' = \Gamma \cup \{A \vee B\} \setminus \{A\}$ , Lemma 8 demonstrates that the conditional is



trivially satisfied. The fourth case is the inductive case. This follows straightforwardly from the transitivity of classical consequence. <sup>M</sup>finish proof ■

Lemma 8 holds that a disjunctive candidate is never sturdy when it must compete against both of its disjuncts. Unfortunately, without a special constraint, disjunctive competitors would also block their disjuncts from obtaining sturdiness.

**Fact 1.**  $\bigcup \neg(\{\{A \vee B\}\} \parallel \{A\}) \in \mathcal{D}(\neg A)$ .

*Proof.* It suffices to show that  $\bigcup \neg(\{\{A \vee B\}\} \parallel \{A\}) = \{\neg A \wedge \neg B\}$  and  $\neg A \wedge \neg B \in \mathcal{D}(\neg A)$ . ■

In order to prevent this unwanted result, we deploy Disjunction Deletion (Definition 18) in the formulation of *sturdy*. Thus, as the following theorem states, a candidate explanans is never forced to compete against a set that contains its disjunction with an arbitrary formula.

**Proposition 4.** If the sequent  $\bigcup \neg(\lfloor S^\Delta \rfloor \parallel \Gamma \cup \{A\}), \Sigma \mid \Gamma, A \mid_{\Theta \cup \Gamma \cup \{A\} \cup \Delta} \Delta$  is the premise in an application of *sturdy*, then  $A \vee B \notin \bigcup \lfloor S^\Delta \rfloor \parallel \Gamma \cup \{A\}$  for any formula  $B$ .

*Proof.* It suffices to note that according to Definition 18, if  $A \in \bigcup S^\Delta$ , then  $A \vee B \notin \bigcup \lfloor S^\Delta \rfloor$  for any arbitrary formula  $B$ . ■

The provisos on *sturdy* are intended to prevent applications of the rule in cases where there is no genuine comparison between a candidate explanans and its competitors—e.g. if a candidate explanans were to be smuggled into the background set of a competitor ( $\Gamma \subseteq \Sigma_i$ ) or vice versa ( $\Omega_i \subseteq \Sigma$ ). Similarly, the requirement that the background set of the candidate explanans form a subset of those of its competitors ( $\Sigma \subseteq \Sigma_i$ ) provides a common set of assumptions against which comparisons can be made. Thus, these provisos provide a level playing field on which candidate explanans may compete for sturdiness.

In combination with Definition 16, this last constraint enables the set of competitors to be culled. For instance, one can prevent an antecedent,  $\Omega_i$ , from belonging to  $S^\Delta$  by constructing a background set for the candidate that includes information that defeats the inference from  $\Omega_i$  to  $\Delta$ . This procedure of culling the competitor set describes how a reasoner goes about holding certain pieces of information (e.g. actual causes) “fixed”.

We turn now to the *abduct.* rule. As the name suggests, this rule is intended to capture the *detachability* of the premises of explanatory arguments. Roughly put, *abduct.* says that if  $\Gamma, A$  is an explanatory argument for  $\Delta$ , and  $\Delta$  is the non-trivial consequence of  $\Gamma'$ , then together,  $\Gamma$  and  $\Gamma'$  license the inference to  $A$ . Since  $A$  only forms part of the explanatory argument for  $\Delta$ , we ought to read *abduct.* as licensing the detachment of a *partial* explanation of  $\Delta$ . The formulation of the rule thus makes detachability a manifest property of  $\blacktriangleright$ -sequents.

There are, however, some peculiarities to the rule that deserve discussion. First, the explanandum ( $\Delta$ ) disappears from the conclusion, leaving only the partial explanans. This feature accords with our desire to present IBE in the strongest form possible. If IBE only licensed inferences to best explanations *or* their explananda, its legitimacy would hardly have roused debate—though its utility might have. Unfortunately, the absence of the explananda in the succedent of *abduct.*'s conclusion means that information is lost in any derivation that contains an application of the rule. The effect, like that of proofs that employ cut, is the failure of analyticity—i.e. some derivations will contain formulas that are not subformulas of those in the end-sequent. The loss of the subformula property is not all that surprising given

the standard characterization of IBE as a form of ampliative inference. We are reassured by the fact that despite this loss, the cut-elimination theorem holds for LEA (Theorem 1).

Second, the defeater set attached to the second premise indicates that the inference that entitles us to the explanandum must be non-trivial. This restriction is justified on the grounds that a tautology should never count as evidence for an explanandum's obtaining. Third, the proviso on *abduct*. prevents the antecedents of the premises from overlapping. This constraint follows from the idea that abductive inferences are only licensed when one has evidence for the explanandum that is independent of the explanans. Note that this means that the second premise of *abduct*. will contain a sequent whose antecedent may have 'competed' with the explanans for sturdiness. This is as it should be, since the competitors include not only potential explanations, but also non-explanatory, evidential inferences. <sup>JM</sup>Should I elaborate on this? Finally, in addition to appearing in the succedent of the conclusion, the (partial) explanans also appears in the background set. This feature is consistent with our understanding of these sets as keeping track of formulas that have shifted from the left to the right of the turnstile.

The logical rules of  $LE^{\blacktriangleright}$  are just the left and right rules for conjunction in  $LK^{\ominus}$ , *modulo*  $\blacktriangleright$ -sequents. We have chosen to restrict the logical operations permitted on  $\blacktriangleright$ -sequents to these out of an abundance of caution. To our ears, inferences from conjunctive explanantia and to conjunctive explananda sound far more natural than those involving disjunctive or negated explanantia/explananda. <sup>JM</sup>Should we say more?

Now that we have in place its constituent systems, let us reflect on the global properties of LEA. Perhaps most striking is the absence of classical structural rules for  $\blacktriangleright$ -sequents. Neither weakening nor cut is permitted. If the antecedents of  $\blacktriangleright$ -sequents could be arbitrarily weakened, then the minimality condition, which *sturdy* secures, would be compromised. Right weakening, on the other hand, does not directly conflict with the properties of explanatory arguments; rather, we deny it because the weakening of explananda by arbitrary disjuncts appears to us as both unnatural and unreasonable. Finally, the absence of cut follows from the fact that our target vocabulary is that which expresses the notion of *immediate* explanation.

**Proposition 5.** It is possible that  $\Sigma \mid \Gamma \mid_{\ominus}^{\blacktriangleright} A, \Delta$  and  $\Sigma' \mid \Gamma', A \mid_{\Psi}^{\blacktriangleright} \Delta'$  are theorems of LEA, but  $\Sigma', \Sigma \mid \Gamma', \Gamma \mid_{\ominus \cup \Psi}^{\blacktriangleright} \Delta, \Delta'$  is not.

*Proof.* Follows from the fact that there is no cut rule for  $\blacktriangleright$ -sequents. ■

While there is no *cut* $\blacktriangleright$  rule, proofs in LEA contain an application of *cut* that was not available in  $LK^{\ominus}$ , namely, one where the cut formula appears in the succedent that follows the application of *abduct*. Fortunately, the cut-elimination theorem can be extended to cover these cases.

**Theorem 1.** Any sequent which is provable in LEA has a cut-free proof.

*Proof.* Lemma 2 gives us cut-elimination for the proofs in  $LK^{\ominus}$ . The following reduction covers the one application of cut that appears in proofs of LEA that does not appear in proofs of  $LK^{\ominus}$ .

$$\begin{array}{c}
\frac{\frac{\bigcup \neg([S^\Delta] \parallel T_\Gamma), \Sigma \mid \Gamma, A \mid_{\Theta \bar{\Gamma} \cup \Delta} \Delta}{\Sigma \mid \Gamma, A \mid_{\Theta \bar{\Gamma} \cup \Delta} \Delta} \text{ sturdy} \quad \frac{\pi}{\vdots} \quad \frac{\Sigma'' \mid \Gamma'' \mid_{\Psi \bar{\Gamma}'' \cup \Delta} \Delta}{\text{abduct.}}}{\frac{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \mid_{\Theta \bar{\Gamma} \cup \Delta \cup \Psi \bar{\Gamma}''} A \quad \Sigma \mid \Gamma, A \mid_{\Theta} \Delta}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \mid_{\Theta \bar{\Gamma} \cup \Delta \cup \Psi \bar{\Gamma}''} \Delta} \text{ cut}} \\
\\
\downarrow \\
\frac{\pi}{\vdots} \quad \frac{\Sigma \mid \Gamma'' \mid_{\Psi} \Delta}{\text{LW}} \quad \frac{\Sigma \mid \Gamma'', \Gamma \mid_{\Psi} \Delta}{\text{BE}} \quad \frac{\Sigma, A \mid \Gamma'', \Gamma \mid_{\Psi} \Delta}{\text{BE}} \quad \frac{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \mid_{\Psi} \Delta}{\text{DE}} \quad \frac{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \mid_{\Delta \cup \Psi} \Delta}{\text{DE}} \quad \frac{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \mid_{\bar{\Gamma} \cup \Delta \cup \Psi} \Delta}{\text{DE}} \quad \frac{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \mid_{\Theta \bar{\Gamma} \cup \Delta \cup \Psi} \Delta}{\text{DE}} \quad \frac{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \mid_{\Theta \bar{\Gamma} \cup \Delta \cup \Psi \bar{\Gamma}''} \Delta}{\text{DE}}
\end{array}$$

■

There are two facts about the proof of Theorem 1 that are of particular significance. First, in contrast with standard normalization procedures for LK, the normalized proof above does not permute the application of cut upwards. Indeed, there is no application of cut whatsoever. Thus, unlike the normalization that secures cut-elimination in  $LK^\Theta$ , there is no need here to guarantee cut-free proofs of the end-sequents of cut-laden paraproofs. The relaxation of this demand is a welcome result, not least because the *sturdy*-rule fails to preserve undefeatedness upwards. (We discuss this point in the conclusion).

Second, there are no  $\blacktriangleright$ -sequents in the normalized proof. While applications of cut in cut-free calculi always involve a ‘detour’ through unnecessary steps, in this case the application of *cut* to the conclusion of *abduct.* renders the prior application of the *sturdy*-rule, and hence the deployment of the  $\blacktriangleright$ -subsystem, particularly gratuitous. Since the normalized proof is a non-branching tree whose leaf is a non-trivial inference—indeed the very inference whose role in *abduct.* is to provide non-explanatory evidence that the ‘explanandum’ obtains—we have in the reduced proof an instance of reasoning that proceeds through explanatory arguments with no epistemic gain. <sup>JM</sup>[Are there broader implications that could drawn out here?](#)

Finally, we can now see that the theorems of LEA constructed with  $\mid_{\blacktriangleright}$  exhibit all of the properties associated with explanatory arguments.

**Theorem 2.** The  $\blacktriangleright$ -sequents that are theorems of LEA are (1) defeasible, (2) minimal, (3) premise-consistent, (4) irreflexive, (5) sturdy, and (6) detachable.

The rules of  $LK^\Theta$  apply to all the formulas in  $\mathcal{L}_\blacktriangleright$ .

The rules of  $LE^\blacktriangleright$  apply to all those formulas in the fragment  $\mathcal{L} \cap \mathcal{L}_\blacktriangleright$ .

**Rules for  $\blacktriangleright$**

$$\frac{\Sigma \mid \Gamma, A \mid_{\Theta} B, \Delta \quad \Sigma' \mid \Gamma' \mid_{\Psi \cup \Gamma' \cup \Delta} B, \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma, A \blacktriangleright B \mid_{\Theta \cup \Psi \cup \Gamma' \cup \Delta} A} \blacktriangleright \vdash^* \quad \frac{\Sigma \mid \Gamma, A \mid_{\Theta} B, \Delta}{\Sigma, A \mid \Gamma \mid_{\Theta} A \blacktriangleright B, \Delta} \vdash \blacktriangleright$$

\* Provided that  $\Gamma \cap \Gamma' \cap \{A\} = \emptyset$ .

**Figure 3:** Rules of  $LEA^+$ .

*Proof.* (1) Defeasibility follows from Definition 11.

(2) Minimality follows from Lemma 11.

(3) Premise-consistency follows from Lemma 3 and Corollary 4.

(4) Irreflexivity follows from Lemma 5 and Corollary 6.

(5) Sturdiness follows from the *sturdy* rule.

(6) Detachability follows from the *abduct.* rule. ■

### 4.3 The extension $LEA^+$

We shall now demonstrate how the system LEA and the language  $\mathcal{L}$  over which it is defined may be extended to include an object-language expression for *best explains why*. We begin with the syntax of the extended language  $\mathcal{L}_\blacktriangleright$ .

**Definition 20** (Syntax of  $\mathcal{L}_\blacktriangleright$ ).

- (1) If  $A \in \mathcal{L}$  then  $A \in \mathcal{L}_\blacktriangleright$ .
- (2) If  $A, B \in \mathcal{L}$  then  $A \blacktriangleright B \in \mathcal{L}_\blacktriangleright$ .

The expressions ‘ $A \blacktriangleright B$ ’ is intended to be read as ‘ $A$  best explains why  $B$ .’ Note that the syntactic definition for  $\blacktriangleright$  is not recursive with respect to  $\mathcal{L}_\blacktriangleright$ . Consequently, the operator  $\blacktriangleright$  is non-iterative. We impose this syntactic constraint on the grounds that the “... best explains why...” locution does not appear to iterate in natural languages—at least not in English.

The rules in Figure 3 define the calculus  $LEA^+$  over  $\mathcal{L}_\blacktriangleright$ . While the rules of  $LK^\Theta$  apply to all the formulas in  $\mathcal{L}_\blacktriangleright$ , the rules for  $LE^\blacktriangleright$  are restricted to the fragment  $\mathcal{L} \cap \mathcal{L}_\blacktriangleright$ . This restriction is intended to prevent the generation of ill-formed formulas along a derivation, e.g.  $A \blacktriangleright (B \blacktriangleright C)$ .

As promised, the extension  $LEA^+$  provides introduction (right) and elimination (left) rules for the *best-explains-why* operator, i.e.  $\blacktriangleright$ . The  $\blacktriangleright \vdash$  rule ought to be familiar—it is essentially the *abduct.* rule with the abducible formula joined to a member of the succedent of the second premise (i.e.  $B$ ) by the  $\blacktriangleright$ -connective and appended to the antecedent of the conclusion. In fact, the  $\blacktriangleright \vdash$  rule is derivable from *abduct.* and LW.

**Proposition 6.**  $\blacktriangleright \vdash$  is a derivable rule in  $LEA^+$ .

*Proof.*

$$\frac{\frac{\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} \Delta \quad \Sigma' \mid \Gamma' \mid_{\Psi \cup \Gamma' \cup \Delta}^{\overline{\phantom{A}}} \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \mid_{\Theta \cup \Psi \cup \Gamma' \cup \Delta}^{\overline{\phantom{A}}} A} \text{abduct.}}{\Sigma', \Sigma, A \mid \Gamma', \Gamma, A \triangleright B \mid_{\Theta \cup \Psi \cup \Gamma' \cup \Delta}^{\overline{\phantom{A}}} A} \text{LW}$$

■

<sup>JM</sup>Is this proposition worth having?

The  $\vdash \triangleright$  rule, on the other hand, represents something quite novel. While it resembles the right rule for  $\rightarrow$  in LK, it is distinguished by two features. First, the active formula in the antecedent of the premise occurs in the background set of the conclusion. This peculiarity is justified by the need to preserve undefeatedness upwards, much in the way that  $\neg \vdash$  does. Second, while the premise is a  $\triangleright$ -sequent, the conclusion is not. This feature captures the sense in which formulas whose main operator is  $\triangleright$  are *explicitly* explanatory claims. As such, these claims can enter into reasoning patterns that do not consist in the making of explanatory arguments, and thus they belong to the class of sequents whose turnstile is unadorned by  $\triangleright$ .

We are now in a position to make good on our promise to provide an expressivist treatment of explanatory vocabulary. Since the deduction theorem serves as the model for logical expressivist theses, it is incumbent upon us to show that a similar theorem holds in  $\text{LEA}^+$ . In order to do so, we must prove the invertibility of  $\vdash \triangleright$ . As we noted above, the fact that latter invokes two distinct classes of consequence relations means that, upon proof of invertibility, the resultant theorem will say something different. Namely, that an expression used in one logic (that of the unadorned turnstile,  $\mid_{\Theta}^{\phantom{\triangleright}}$ ) encodes the rules of another logic ( $\mid_{\Theta}^{\triangleright}$ ). To distinguish this claim from the standard deduction theorem, we refer to it as a *quasi-deduction theorem*.

**Theorem 3** (Quasi-Deduction Theorem).  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} B, \Delta$  is provable in  $\text{LEA}^+$  if and only if  $\Sigma, A \mid \Gamma \mid_{\Theta}^{\phantom{\triangleright}} A \triangleright B, \Delta$  is.

*Proof.*  $(\Rightarrow)$  Follows from  $\vdash \triangleright$ .

$(\Leftarrow)$  By induction on proof-height. **Base Case:** If  $\Sigma, A \mid \Gamma \mid_{\Theta}^{\phantom{\triangleright}} A \triangleright B, \Delta$  is an axiom, then, since  $A \triangleright B$  is not atomic,  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} B, \Delta$  would have to be an axiom. However, there are no axioms with  $\triangleright$ -sequents. Rather, all such sequents are derived via *sturdy*. Thus,  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} B, \Delta$  follows by *sturdy*.

**Inductive Step:** Assume inversion up to height  $n$  and let  $\Sigma, A \mid \Gamma \mid_{\Theta}^{\phantom{\triangleright}} A \triangleright B, \Delta$  be the root of a proof of height  $n+1$ . There are two cases:

Case 1: If  $A \triangleright B$  is not principal in the last rule, it has one or two premises,  $\Sigma', [A] \mid \Gamma' \mid_{\Theta}^{\phantom{\triangleright}} A \triangleright B, \Delta'$  and  $\Sigma'', [A] \mid \Gamma'' \mid_{\Theta}^{\phantom{\triangleright}} A \triangleright B, \Delta''$  of derivation height  $\leq n$ , where  $[A]$  indicates that  $A$  is in at least one of the premises (if there are two). By inductive hypothesis  $\Sigma' \mid \Gamma', A \mid_{\Theta}^{\triangleright} B, \Delta'$  and  $\Sigma'' \mid \Gamma'', A \mid_{\Theta}^{\triangleright} B, \Delta''$  are obtained by a proof of height  $n$ . Now apply the last rule to obtain  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} B, \Delta$  with a proof of height  $\leq n+1$ .

Case 2: If  $A \triangleright B$  is principal in the last rule, the premise  $\Sigma \mid \Gamma, A \mid_{\Theta}^{\triangleright} B, \Delta$  has a proof of height  $\leq n$ . ■

<sup>JM</sup>Discuss quasi-deduction theorem and show how it captures the spirit of logical expressivism.

We can now show that  $\text{LEA}^+$  is a conservative extension of LEA. First, we prove that the cut elimination theorem holds for  $\text{LEA}^+$ .

**Theorem 4.** Any sequent provable in  $\text{LEA}^+$  has a cut-free proof.

*Proof.* Since Theorem 2 establishes cut elimination for LEA, we need to show that every application of cut that occurs in proofs of  $\text{LEA}^+$  but not in proofs of LEA is eliminable. There is only one such application, namely, that which cuts the formula  $A \blacktriangleright B$  from the conclusions of  $\blacktriangleright \vdash$  and  $\vdash \blacktriangleright$ . The following reduction covers this application.

$$\begin{array}{c}
\begin{array}{c} \pi_1 \\ \vdots \end{array} \quad \begin{array}{c} \pi_2 \\ \vdots \end{array} \\
\frac{\Sigma \mid \Gamma, A \mid_{\Theta}^{\blacktriangleright} B, \Delta \quad \Sigma' \mid \Gamma' \mid_{\Psi \cup \overline{\Gamma'} \cup \Delta} B, \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma, A \blacktriangleright B \mid_{\Theta \cup \Psi \cup \overline{\Gamma'} \cup \Delta} A} \blacktriangleright \vdash \quad \frac{\Sigma \mid \Gamma, A \mid_{\Theta}^{\blacktriangleright} B, \Delta}{\Sigma, A \mid \Gamma \mid_{\Theta} A \blacktriangleright B, \Delta} \vdash \blacktriangleright \\
\hline
\Sigma', \Sigma, A \mid \Gamma', \Gamma \mid_{\Theta \cup \Psi \cup \overline{\Gamma'} \cup \Delta} A, \Delta \quad \text{cut} \\
\downarrow \\
\begin{array}{c} \pi_1 \\ \vdots \end{array} \quad \begin{array}{c} \pi_2 \\ \vdots \end{array} \\
\frac{\Sigma \mid \Gamma, A \mid_{\Theta}^{\blacktriangleright} B, \Delta \quad \Sigma' \mid \Gamma' \mid_{\Psi \cup \overline{\Gamma'} \cup \Delta} B, \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \mid_{\Theta \cup \Psi \cup \overline{\Gamma'} \cup \Delta} A} \text{abduct.} \\
\hline
\Sigma', \Sigma, A \mid \Gamma', \Gamma \mid_{\Theta \cup \Psi \cup \overline{\Gamma'} \cup \Delta} A, \Delta \quad \text{RW}
\end{array}$$

■

The conservativity of  $\text{LEA}^+$  follows immediately from cut's eliminability.

**Corollary 4.1.** Every theorem in  $\text{LEA}^+$  that only contains formulas from  $\mathcal{L}$  is a theorem of LEA.

*Proof.* Since  $\blacktriangleright \vdash$  and  $\vdash \blacktriangleright$  are the only rules in  $\text{LEA}^+$  that are not in LEA, the only source of new theorems formulated in  $\mathcal{L}$  are those derived via an application of cut to the conclusions of these two rules. It follows from Theorem 4 that this application of cut is eliminable. ■

## 5 Conclusion

In this paper, we have argued that there is a viable inferentialist-expressivist treatment of explanatory vocabulary. More specifically, we have shown how explanatory arguments and IBE can be represented by a sequent calculus and how that calculus can be conservatively extended to a language that contains explicitly explanatory expressions. We conclude by discussing some of the peculiarities and limitations of the current approach as well as the prospects for future work.

Since our calculi are constructed on the basis of LK, our defeasible sequents represent relations among sets of formulas, while our connectives are operations on formulas. In LK this discrepancy between the relata of consequence relations and the relata of connectives is completely natural. However, because our  $\blacktriangleright$ -sequents are intended to capture an exhaustive relation—namely, the relation of loveliest potential *exhaustive* immediate explanation—the fact that the  $\blacktriangleright$  connective only holds among formulas might threaten a change in meaning.

Fortunately, this concern is easily allayed. What  $\vdash \blacktriangleright$  says is that if  $\Gamma, A$  is the best *exhaustive* immediate explanation of  $B$ , then  $A$  is the best *exhaustive* immediate explanation of  $B$  given (all of)  $\Gamma$ . This reading of the rule corresponds nicely to the pragmatics of explanation. We are rarely in search of *complete* explanations. Rather, our inquiries aim at *the* (best) explanation of some phenomena. The latter may be thought of as a *selection* from the former. To proffer a *selected* explanation is to put forth one claim as an explanans while *assuming* that the remaining components of an exhaustive explanation hold. For instance, to explain why George ordered the chocolate cake, we might appeal to a stable preference—e.g. “Chocolate cake is George’s favorite dessert.” But the sufficiency of this explanation rests on a number of assumptions—for instance, that a preference for chocolate cake would lead someone to choose chocolate over carrot cake. Our formalism does justice to this fact: someone entitled to an exhaustive explanatory argument is entitled to claim that some element of the premises ( $A$ ) is the best explanation of the conclusion ( $B$ ), so long as one is entitled to the remaining premises ( $\Gamma$ ). In fact, our insistence that defeasible sequents be paired with background sets ( $\Sigma$ ) supports the view that even the propriety of exhaustive explanatory arguments depends upon what background assumptions are in play.

Of course, what makes one selected explanation acceptable in practice as opposed to another is no doubt in part a function of speaker interests. Consequently, a satisfactory theory of best explanations must account for the role that the practical commitments of speakers (i.e. plans, projects, preferences, interests) play in determining *which* components of an exhaustive explanation can serve as the best explanation. To do so, the theory must provide a means for conceptualizing the pragmatics of explanation.

### Wish List

1. Our calculi provide opportunities for the formalization of different target vocabularies. For instance, while our aim was to offer an expressivist treatment of immediate explanation, there is no reason to think that a similar treatment of *mediate* explanation is unavailable. Such an account would need to permit a version of cut in LEA. Adding the cut rule from  $LK^\ominus$  would mean that transitivity can fail when one of the sequents with the cut formula is defeated. Is this just the sort of non-transitivity we want?
2. Unlike many nonmonotonic logics, cautious monotonicity does not hold in LEA. How would we include this rule? Would we want to? Think about Ulf’s line on this: CM says that you can always add implicit content *explicitly* to the premises.
3. What additional logical rules should be included in  $LE^\blacktriangleright$ ?
4. How to represent probabilistic inferences? Treat defeater sets as things that would bring the inference below a threshold probability.

## References

- Brandom, Robert. 2008. *Between Saying and Doing : Towards an Analytic Pragmatism*. Oxford ; New York: Oxford University Press.
- Hempel, Carl G. 1965. *Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science*. New York: Free Press.
- Hlobil, Ulf. 2016. “A Nonmonotonic Sequent Calculus for Inferentialist Expressivists”. In *Logica Yearbook 2015*, ed. by Pavel Arazim and Michal Dančák, 87–105. London: College Publications.
- Lange, Marc. 2009. “Why Do the Laws Explain Why?” In *Dispositions and Causes*, ed. by T. Handfield. Mind Association Occasional Series. Oxford University Press.

- Lipton, P. 2004. *Inference to the Best Explanation*. International library of philosophy and scientific method. Routledge.
- Mitchell, Sandra D. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge University Press.
- Piazza, Mario, and Gabriele Pulcini. 2015. “Unifying logics via context-sensitiveness”. *Journal of Logic and Computation*.
- . 2016. “Uniqueness of axiomatic extensions of cut-free classical propositional logic”. *Logic Journal of IGPL* 24 (5): 708–718.
- Ripley, David. 2011. “Paradoxes and Failures of Cut”. *Australasian journal of philosophy* 91 (1): 139–164.
- Skyrms, Brian. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven, CT: Yale University Press.
- Tennant, Neil. 2014. “Natural Deduction and Sequent Calculus for Intuitionistic Relevant Logic”. *The Journal of Symbolic Logic* 52 (03): 665–680.
- Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. vi, 410. New York: Oxford University Press.