A Logic of Best Explanations

April 12, 2017

1 Introduction

The past two decades have witnessed a renewed interest among logicians in explanatory reasoning. The legacy of logical positivism and its insistence that there was no logic of scientific discovery has at last given way to ever more sophisticated attempts to formalize *abduction* or *inference to the best explanation* (IBE). 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*) and that of evaluating hypotheses (i.e. *inference to the best explanation*). Most philosophers of science concede that what constitutes the 'best explanation' is a function of contextual considerations. At the very least the best explanation will be partially determined by what plausible explanations are available. (This line of thought has lead some, such as Bas van Fraassen to see IBE as wholly unreliable). In any case, if the notion of 'best explanation' contains a contextual parameter, then any formalization of IBE will need to reserve a role for extra-logical criteria. While there have been some attempts to meet this challenge by incorporating preference rankings constrained by structural properties, the apparent context-sensitivity of best explanations has lead most logicians to focus on the first of the two patterns of abduction.

Unfortunately, even among the treatments of hypothesis generation, or what Atocha Aliseda calls *plain abduction*, there has been little effort to represent the uniquely *explanatory* character of the premises—namely, the behavior of the explanatory connective(s) expressed by English locutions such as "That…(best) explains why…" or "…because …". Instead, abduction is all too often treated as a kind of backwards Modus Ponens with material implication standing in for the explanatory connective. As a result, the relation between explanans and explanandum is reduced to the classical consequence relation or to one of its restricted forms.[1]

Logicians' hesitancy to formalize explanatory connectives is not surprising. Besides the supposed context-sensitivity of "*best* explains," explanatory relations in general exhibit a challenging cluster of properties. They appear to be contradiction-intolerant, (possibly) non-transitive, (probably) asymmetric, (definitely) non-monotonic, and (worst of all) irreflexive. Some of these properties are captured by well-established nonclassical logics, but others, such as irreflexivity, run counter to even the most heterodox conceptions of logical consequence.

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 ultra-weak system for conditional logic. The conditional they use, that of Brian Chellas's **CR**, blocks the validity

[1] A notable exception to this trend is the work of adaptive logicians...

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. The absence of such a rule is by no means problematic for Beirlaen and Alesida's account, whose aim is to provide a model of abduction in which the conditional of **CR**, rather than material implication, appears in the major premise. But this lacuna

link up explanative connective and consequence relation

Why would we want to do so?

2 Explanatory Vocabulary

To start, let us specify our target. We aim to show that the role that a particular class of expressions plays in a language is that of enabling speakers to express their commitment to certain inferences, rather than to describe features of the world. The class we have in mind includes those expressions which are given in English as "That…(best, possibly, actually) explains why…," and its nominalization "That…is the (best, possible, actual) 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 the pattern of inference dubbed "Inference to the Best Explanation" (IBE). In its simplest rendition, IBE takes the following form:

$$\begin{array}{l} A \text{ best explains why } B \\ \underline{B \qquad\qquad\qquad\qquad} \\ \therefore A \end{array}$$

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." (**Lipton2004**). For this reason, Lipton suggests that the best explanation ought to be construed as the best *potential* explanation.[2] 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

[2] 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.

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*.

JM I just built non-transitivity into our target. This seems cleaner to me. Now it's no longer a property we have to justify.

3 Properties of Explanatory Arguments

If, *ex hypothesi*, explanatory expressions *do* serve to make explicit certain inferential commitments, the first question that arises is, what kinds of inferences are being made explicit? Let us call these target phenomena *explanatory arguments*. Our question then becomes: what are the properties of explanatory arguments that distinguish them from other inferences? We can certainly reconstruct a Hempelian answer to this question. For Hempel, (certain) explanatory arguments are just inferences with true premises that include at least one lawlike statement. However, given the notorious difficulties that plagued Hempel's deductive-nomological account, we should resist cleaving too closely to this model. Instead, we propose to work backwards from scientific practice concerning the appropriate use of explanatory expressions and the general inferential relationships they exhibit. In other words, we proceed from the assumption that the properties of explanatory arguments may be read off of the behavior of explanatory practices.

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 explanatory arguments in question are defeated. Since classical deductive inferences are monontonic—i.e. the premises of a valid inference can be arbitrarily expanded and the resulting inference is also valid—capturing the defeasibility of explanatory arguments cannot be done within a strictly deductive system.

Second, explanatory arguments are *minimal* in the sense that all of their premises are needed to infer their conclusion. More precisely, minimality insists that (M1) no proper subset of the premises of an explanatory argument are explanatory (of the same explanandum), (M2) nor any set comprised of conjunctions of members of a proper subset of the premises, (M3) nor any set comprised of disjunctions of two or more members of a subset of the premises . For example, if the set of sentences $\{A, B, C\}$ is a minimal explanatory argument for D, then none of the following are explanatory arguments for D: $\{A, B\}$, $\{A \wedge B\}$, $\{B \wedge C\}$, $\{A \vee B\}$, $\{A \vee (B \vee C)\}$. 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. The minimality of explanatory arguments is thus another reason why they cannot be modeled by deductive, i.e. monotonic inferences.

Third, since contradictions do not explain, explanatory arguments must have consistent premises. Classically valid arguments adhere to the principle of *ex contradictione quodlibet* (ECQ)—i.e. $A, \neg A \vdash B$ for any arbitrary B—meaning that any argument with inconsistent premises is classically valid. Logical systems that renounce ECQ are called either paraconsistent or premise-consistent. Paraconsistent systems are able to represent reasoning that is *tolerant* of inconsistency and that may, in some cases, even validate contradictions. By contrast, a *premise-consistent* logic aims, in the opposite direction, to represent reasoning that is *intolerant* of inconsistency. Since contradictions never explain, explanatory arguments are premise-consistent, not paraconsistent. Excluding theorems with contradictory premises, however, comes at the cost of logical strength. In our system, for example, disjunctive syllogism is not a theorem, so it can only be retained as a rule. Finally, note that premise consistency describes yet another form of nonmonotonic behavior.

Fourth, explanatory arguments are *irreflexive*.[3] For instance, "the litmus paper turned red because it turned it 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.[4] One particular difficulty arises from the fact that irreflexive logics cannot easily recover Modus Ponens (MP), yet many explanations require Modus Ponens. Our irreflexive system preserves Modus Ponens.

Fifth, explanations are *stable*, which philosophers of science have analyzed in different ways (**Hempel1965**; **Lange2009**; **Mitchell2003**; **Skyrms1980**; **Woodward2003**). 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, as inferentialists, we take explanatory arguments to be the fundamental bearers of stability. In what follows, we develop a distinctively inferentialist brand of stability that we call "sturdiness."

We baptize as *non-trivial* those defeasible inferences that are premise-consistent and irreflexive. Sturdiness is then 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

[3] While a *non-reflexive* system would be one in which the principle of reflexivity fails, perhaps only on occasion, an *irreflexive* system is one in which no interface of reflexivity holds.

[4] 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.

[0] Logicians of inferentialist and proof-theoretic persuasion have already explored systems in which transitivity fails (**Ripley2011**; **Tennant2014**; **Hlobil2016**),

explanans, we may make further predictions, design interventions, and construct new models. This is the animating idea behind Inference to the Best Explanation. A consequence relation that captures explanation must thereby underwrite abduction.

$^{\mathrm{JM}}$Transition

4 The Logic of Explanatory Arguments and IBE

In this section, we develop sequent calculi for explanatory arguments and the vocabulary that makes them explicit. 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, $\mathsf{LK}^{\ominus}$, is based on a variant of Gentzen's sequent calculus for classical logic that was proposed by **Piazza2015** 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 **Piazza2015** introduced their calculus, we retain the key technical features of $\mathsf{LK}^{\mathcal{S}}$ and inherit their proof of the cut-elimination theorem. Our modifications are intended, on one 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, $\mathsf{LE}^{\blacktriangleright}$ introduces an additional class of sequents and rules that govern their interaction with those of $\mathsf{LK}^{\ominus}$, 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 function of locutions like *best explains why* is to express commitment to 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 the locution *A best explains why B* does make explicit a commitment to an explanatory argument. 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 *best explains why* B*)* $\wedge$ B*] 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.[5] 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 $\mathsf{LK}^{\ominus}$ is intended to represent those sets of inferences to which both IBE and *candidate* explanatory arguments belong. The class of consequence relations introduced by $\mathsf{LE}^{\blacktriangleright}$ 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

[5] 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.

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 $At = \{p_1, \ldots, p_n, q_1, \ldots, q_n\}$, 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 $\mathbf{S}, \mathbf{T}, \mathbf{U}$ 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, Θ, called a *defeater set* and to the far left of the turnstile we add another, Σ, called a *background set*.

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

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 $\mathsf{LK}^{\mathcal{S}}$ developed by **Piazza2015**. 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, π, constructed by recursive applications of the rules (excluding *cut*), the following holds: if a defeated sequent occurs in a branch of π, 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.[6]

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

[6] In **Piazza2015** 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.

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 **Piazza2015** to the extent that proofs in LEA preserve both validity and undefeatedness. However, since the rules in our calculi, and those of $\mathsf{LE}^{\blacktriangleright}$ in particular, are not deductive, it cannot be deductive 'validity' that is preserved. Instead, we follow **Brandom2008** 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)* Γ *is entitled to (at least one member of)* Δ, *given the background of* Σ.

$^{\mathrm{JM}}$Issues with normative pragmatics interpretations of multiple succedent sequents.

$^{\mathrm{JM}}$transition

Definition 1 (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 sequent with a *background set*, Σ, and a *defeater set*, Θ, attached:[7]

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

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

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

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

Definition 2 ($\mathcal{D}$-Rules)**.** Let $\mathcal{D}$ be the following set of rules:

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

$\mathcal{D}_{\vee}$: $A \in \mathcal{L}$ 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.

[7] In general, the respective roles played by *control sets*, *compatibility*, and *soundness* in Piazza and Pulcini's $\mathsf{LK}^{\mathcal{S}}$ 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 **Piazza2015**, 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 6 and Lemma 1 are taken over from **Piazza2015** with very little modification.

Definition 3 ($\mathcal{D}(\Theta)$)**.** Let $\mathcal{D}(\Theta)$ be the closure of Θ under the $\mathcal{D}$-rules as well as the commutativity, associativity, and distributivity of conjunction and disjunction, respectively.

Definition 4 (Compatibility)**.** A set of formulas, Γ, is said to be compatible with a defeater set, Θ, just in case the conjunction of the members of Γ 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 Γ, $\Gamma \succsim \emptyset$.

Definition 5 (Defeat)**.** A defeasible sequent, $\Sigma \mid \Gamma \vdash_{\Theta} \Delta$, 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, Θ, 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 2, 3, and 4. ■

4.1 $\mathsf{LK}^{\ominus}$

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 $\mathsf{LK}^{\ominus}$ in Figure 1. These rules are designed to generate trees that preserve validity *downward* and undefeatedness *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 6 (Proof, Paraproof)**.** For a rooted, finitely branching tree π whose nodes are sequents of $\mathsf{LK}^{\ominus}$ (LEA), and which is recursively built up from axioms by means of the rules of $\mathsf{LK}^{\ominus}$ (LEA), if each sequent in π is undefeated, then π is said to be a proof of $\mathsf{LK}^{\ominus}$ (LEA), otherwise π is called a paraproof.

The axioms of $\mathsf{LK}^{\ominus}$ 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.[8] 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 7 (Constraints on Defeater Sets for Axioms)**.** If $\cdot \mid p \vdash_{\Theta} p$ is a logical axiom and $\Sigma \mid X \vdash_{\Psi} Y$ is a proper axiom, then

(i) Θ and Ψ 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 $\mathsf{LK}^{\ominus}$, the rules of $\mathsf{LE}^{\blacktriangleright}$ will prevent this property from being transfered 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

[8] 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, **Piazza2016** 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-compete*—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 **Piazza2016**). Our proper axioms have been formulated in conformity with this constraint.

Logical Axioms

$$\frac{}{\cdot \mid p \vdash_{\Theta} p}\ \mathit{log.\ ax.}$$

Proper Axioms

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

Cut Rule

$$\frac{\Sigma \mid \Gamma \vdash_{\Theta} A, \Delta \qquad \Sigma' \mid \Gamma', A \vdash_{\Psi} \Delta'}{\Sigma', \Sigma \mid \Gamma', \Gamma \vdash_{\Theta \cup \Psi} \Delta, \Delta'}\ \mathit{cut}$$

Structural Rules

$$\frac{\Sigma \mid \Gamma \vdash_{\Theta} \Delta}{\Sigma \mid \Gamma, A \vdash_{\Theta} \Delta}\ \mathsf{LW} \qquad\qquad \frac{\Sigma \mid \Gamma \vdash_{\Theta} \Delta}{\Sigma \mid \Gamma \vdash_{\Theta} \Delta, A}\ \mathsf{RW}$$

$$\frac{\Sigma \mid \Gamma \vdash_{\Theta} \Delta}{\Sigma \mid \Gamma \vdash_{\Theta \cup \Psi} \Delta}\ \mathsf{DE} \qquad\qquad \frac{\Sigma \mid \Gamma \vdash_{\Theta} \Delta}{\Sigma, A \mid \Gamma \vdash_{\Theta} \Delta}\ \mathsf{BE}$$

Logical Rules

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

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

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

† Provided that $\{A\} \succsim \Theta$ and $\{B\} \succsim \Psi$.

Figure 1: Rules for LK^{Θ}

of LK^{Θ}. 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 π is a tree whose root is $\Sigma \mid \Gamma \vdash_{\Theta'} \Delta$ and $\cdot \mid p \vdash_{\Theta} p$ occurs in the leaves of π and $A \in \Theta$, then $A \in \Theta'$.

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

Proof. Follows directly from the rules of LK^{Θ}. ■

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.

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.[9] 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^{Θ} 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 (†) 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.[10] 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^{Θ} has a cut-free proof.

[9] 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.

[10] See **Piazza2015** for an example.

Proof. As noted above, the similarity between our $\mathsf{LK}^{\ominus}$ and Piazza et al.'s $\mathsf{LK}^{\mathcal{S}}$ 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 **Piazza2015** Theorem 19. ■

4.2 $\mathsf{LE}^{\blacktriangleright}$ and LEA

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

Definition 8 ($\neg(\Gamma)$)**.** Let $\neg(\Gamma)$ stand for the following operation:

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

The operation easily extends to sets of sets:

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

Definition 9 ($\overline{\Gamma}$)**.** Let $\overline{\Gamma}$ stand for the following set function:

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

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

Definition 10 (Competitor Set, $\mathbf{S}^{\Delta}$)**.** Let $\mathbf{S}^{\Delta}$ stand for the following set:

$$\mathbf{S}^{\Delta} =_{df} \{\Omega_i : \Sigma_i \mid \Omega_i \mid\!\overset{\mathsf{LK}^{\ominus}}{\underset{\Psi_i \cup \overline{\Omega}_i \cup \Delta}{\rule{3em}{0.4pt}}} \Delta\}$$

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

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 $\mathsf{LK}^{\ominus}$ that have Δ 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 11 (Disjunction Deletion, $\lfloor\mathbf{S}\rfloor$)**.**

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

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 12 ($\Delta \mathbin{\backslash\!\backslash} \Gamma$)**.**

$$\Delta \mathbin{\backslash\!\backslash} \Gamma =_{df} \begin{cases} \Delta & \text{if } \Delta \subseteq \Gamma \\ \Delta\backslash\Gamma & \text{otherwise} \end{cases}$$

We extend the definition to sets of sets as follows:

$$\mathbf{S} \mathbin{\backslash\!\backslash} \Gamma =_{df} \{\Delta \mathbin{\backslash\!\backslash} \Gamma : \Delta \in \mathbf{S}\}$$

The set denoted by $\mathbf{S} \setminus\!\setminus \Gamma$ consists of those members of $\mathbf{S}$ that are subsets of Γ and the set-complements relative to Γ of those members that are not subsets of Γ. As we shall see, Definitions 11 and 12 are needed to secure minimality.

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

Structural Rules

$$\dfrac{\Sigma \mid \Gamma \vdash^{\blacktriangleright}_{\Theta} \Delta}{\Sigma \mid \Gamma \vdash^{\blacktriangleright}_{\Theta\cup\Psi} \Delta}\ \mathsf{DE}^{\blacktriangleright} \qquad\qquad \dfrac{\Sigma \mid \Gamma \vdash^{\blacktriangleright}_{\Theta} \Delta}{\Sigma, A \mid \Gamma \vdash^{\blacktriangleright}_{\Theta} \Delta}\ \mathsf{BE}^{\blacktriangleright}$$

Mixed Rules

$$\dfrac{\bigcup\neg(\lfloor\mathbf{S}^{\Delta}\rfloor \setminus\!\setminus \Gamma),\, \Sigma \mid \Gamma \vdash_{\Theta\cup\overline{\Gamma}\cup\Delta} \Delta}{\Sigma \mid \Gamma \vdash^{\blacktriangleright}_{\Theta\cup\overline{\Gamma}\cup\Delta} \Delta}\ \mathit{sturdy}^{\ddagger} \qquad\qquad \dfrac{\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} \Delta \qquad \Sigma' \mid \Gamma' \vdash_{\Psi\cup\overline{\Gamma'}\cup\Delta} \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \vdash_{\Theta\cup\Psi\cup\overline{\Gamma'}\cup\Delta} A}\ \mathit{abduct.}$$

Logical Rules

$$\dfrac{\Sigma \mid \Gamma, A, B \vdash^{\blacktriangleright}_{\Theta} \Delta}{\Sigma \mid \Gamma, A \wedge B \vdash^{\blacktriangleright}_{\Theta} \Delta}\ \wedge\vdash^{\blacktriangleright} \qquad\qquad \dfrac{\Sigma \mid \Gamma \vdash^{\blacktriangleright}_{\Theta} A, \Delta \qquad \Sigma' \mid \Gamma \vdash^{\blacktriangleright}_{\Psi} B, \Delta}{\Sigma', \Sigma \mid \Gamma \vdash^{\blacktriangleright}_{\Theta\cup\Psi} A \wedge B, \Delta}\ \vdash^{\blacktriangleright}\wedge$$

‡ 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 $\mathsf{LE}^{\blacktriangleright}$

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 Δ and its background set includes a set obtained from $\mathbf{S}^{\Delta}$. The second step is captured by the fact that $\bigcup\neg(\lfloor\mathbf{S}^{\Delta}\rfloor \setminus\!\setminus \Gamma)$ 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 Δ, with the caveat that the members of Γ are removed from any antecedent that is a superset of Γ. 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. $\vdash^{\blacktriangleright}$.

By restricting premise sequents to those whose defeater sets contain $\overline{\Gamma}$— where Γ 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 $\Sigma \mid \Gamma, A, \neg A \vdash^{\blacktriangleright}_{\Theta} \Delta$ is not a theorem of LEA.

Proof. Suppose for *reductio* that $\Sigma \mid \Gamma, A, \neg A \vdash^{\blacktriangleright}_{\Theta} \Delta$ 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 $\Gamma, A, \neg A \vdash_{\Theta \cup \overline{\{A, \neg A\}} \cup \Delta} \Delta$. By Definition 9, $\bigwedge\{A, \neg A\} \in \mathcal{D}(\overline{\{A, \neg A\}})$, since $A \wedge \neg A \in \{B \wedge \neg B : B \in \mathcal{L}\}$. Thus, *contra* our supposition, $\Sigma \mid \Gamma, A, \neg A \vdash^{\blacktriangleright}_{\Theta} \Delta$ is defeated. ■

Lemma 4. The sequent $\Sigma \mid \Gamma, A \wedge \neg A \vdash^{\blacktriangleright}_{\Theta} \Delta$ is not a theorem of LEA.

Proof. Same as for Lemma 3. ■

The *sturdy* rule also prevents $\blacktriangleright$-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 $\mathsf{LE}^{\blacktriangleright}$.

Lemma 5. The sequent $\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} A, \Delta$ is not a theorem of LEA.

Proof. Suppose for *reductio* that $\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} A, \Delta$ 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 $\Gamma, A \vdash_{\Theta \cup \{A\} \cup \Delta} A, \Delta$. But, since (by Definition 3) $\bigwedge(\Gamma \cup \{A\}) \in \mathcal{D}(A)$, this sequent is defeated, contradicting our supposition. ■

Lemma 6. The sequent $\Sigma \mid \Gamma, A \wedge B \vdash^{\blacktriangleright}_{\Theta} A, \Delta$ 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. Σ) whenever possible.

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

Proof. Suppose for *reductio* that there is no proof π of $A, \neg A \vee B \vdash_{\Theta \cup \overline{\{A, \neg A \vee B\}} \cup \{B\}} B$ from $A, \neg A \vee B \vdash_{\Theta} B$. It follows that either π is not a paraproof or $A, \neg A \vee B \vdash_{\Theta \cup \overline{\{A, \neg A \vee B\}} \cup \{B\}} B$ is defeated. But $A, \neg A \vee B \vdash_{\Theta \cup \overline{\{A, \neg A \vee B\}} \cup \{B\}} B$ follows from $A, \neg A \vee B \vdash_{\Theta} B$ by a single application of DE, and thus π is at least a paraproof. If $A, \neg A \vee B \vdash_{\Theta \cup \overline{\{A, \neg A \vee B\}} \cup \{B\}} B$ is defeated, then $\bigwedge\{A, \neg A \vee B\} \in \mathcal{D}(\Theta \cup \overline{\{A, \neg A \vee B\}} \cup \{B\})$. But $A \wedge (\neg A \vee B) \notin \mathcal{D}(\{B\})$, and by Definition 3 and 9 it follows that $A \wedge (\neg A \vee B) \notin \mathcal{D}(\overline{\{A, \neg A \vee B\}})$. Therefore, it must be the case that $A \wedge (\neg A \vee B) \in \mathcal{D}(\Theta)$. But by hypothesis $A, \neg A \vee B \vdash_{\Theta} B$ is undefeated. Thus, *contra* our supposition, there is a proof of $A, \neg A \vee B \vdash_{\Theta \cup \overline{\{A, \neg A \vee B\}} \cup \{B\}} B$. ■

Less obvious than the achievement of irreflexivity is the fact that *sturdy* also ensures that $\blacktriangleright$-sequents are minimal in the sense articulated above by conditions (M1)–(M3).

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

Proof. We proceed by proving the contrapositive. Suppose for conditional proof that $\Gamma \vdash^{\blacktriangleright}_{\Theta} \Delta$ is a theorem of LEA. It follows from *sturdy* that $\Gamma \vdash_{\Theta} \Delta$ is a theorem and from Definition 10 that $\Gamma \in \lfloor \mathbf{S}^{\Delta} \rfloor$. Next suppose for *reductio* that $\Gamma, A \vdash^{\blacktriangleright}_{\Theta} \Delta$ is also a theorem. From *sturdy* it follows that $\bigcup \neg(\lfloor \mathbf{S}^{\Delta} \rfloor \setminus\!\!\setminus \Gamma \cup \{A\}) \mid \Gamma, A \vdash_{\Theta \cup \overline{\Gamma \cup \{A\}} \cup \Delta} \Delta$ is a theorem. Since $\Gamma \subseteq \Gamma \cup \{A\}$, it follows from Definition 10 and 12 that $\neg(\Gamma) \subset \bigcup \neg(\lfloor \mathbf{S}^{\Delta} \rfloor \setminus\!\!\setminus \Gamma \cup \{A\})$. But from Definition 3 it follows that $\bigwedge \neg(\Gamma) \in \mathcal{D}(\overline{\Gamma \cup \{A\}})$, so, according to Definition 4 $\bigcup \neg(\lfloor \mathbf{S}^{\Delta} \rfloor \setminus\!\!\setminus \Gamma \cup \{A\}) \mid \Gamma, A \vdash_{\Theta \cup \overline{\Gamma \cup \{A\}} \cup \Delta} \Delta$ is defeated, contradicting our supposition. Thus, if $\Gamma \vdash^{\blacktriangleright}_{\Theta} \Delta$ is a theorem, then $\Gamma, A \vdash^{\blacktriangleright}_{\Theta} \Delta$ is not. It follows by contraposition and double negation that if the sequent $\Gamma, A \vdash^{\blacktriangleright}_{\Theta} \Delta$ is a theorem of LEA, then $\Gamma \vdash^{\blacktriangleright}_{\Theta} \Delta$ is not. ■

Lemma 8. If the sequent $\Sigma \mid A, B, C \vdash^{\blacktriangleright}_{\Theta} \Delta$ is a theorem of LEA, then $\Sigma \mid B \wedge C \vdash^{\blacktriangleright}_{\Theta} \Delta$ is not.

Proof. It suffices to show that $\bigcup \neg(\lfloor \mathbf{S}^{\Delta} \rfloor \setminus\!\!\setminus B \wedge C), \Sigma \mid B \wedge C \vdash^{\blacktriangleright}_{\Theta} \Delta$ is defeated when $\{A, B, C\} \in \lfloor \mathbf{S}^{\Delta} \rfloor$. First note that $\bigwedge \neg(\{A, B, C\}) = \neg A \wedge \neg B \wedge \neg C$ and that $\neg B \vee \neg C \in \mathcal{D}(\overline{B \wedge C})$ and thus, by the $\mathcal{D}_{\vee}$-rule, $\{\neg B, \neg C\} \subset \mathcal{D}(\overline{B \wedge C})$. It follows by the $\mathcal{D}_{\wedge}$-rule that $\neg A \wedge \neg B \wedge \neg C \in \mathcal{D}(\overline{B \wedge C})$, and thus $\bigwedge \neg(\{\{A, B, C\}\}) \in \mathcal{D}(\overline{B \wedge C})$. The rest of the proof mirrors that of Lemma 7 and is left as an exercise for the reader. ■

Lemma 9. If the sequents $\Sigma' \mid A \vdash_{\Theta'} \Delta$ and $\Sigma'' \mid B \vdash_{\Theta''} \Delta$ are theorems of LEA then $\Sigma \mid A \vee B \vdash^{\blacktriangleright}_{\Theta} \Delta$ is not.

Proof. Suppose for *reductio* that $\Sigma, \mid A \vee B \vdash^{\blacktriangleright}_{\Theta} \Delta$ is a theorem. It follows by hypothesis and *sturdy* that $\Sigma, \bigcup \neg(\{\{A, B\}\} \setminus\!\!\setminus \{A \vee B\}) \mid A \vee B \vdash_{\Theta \cup \overline{A \vee B} \cup \Delta} \Delta$ is a theorem. $\{\{A, B\}\} \setminus\!\!\setminus \{A \vee B\} = \{\{A, B\}\}$ and by Definition 8, $\bigcup \neg(\{\{A, B\}\}) = \{\neg A, \neg B\}$. Since $\bigwedge \{\neg A, \neg B\} = \neg A \wedge \neg B$, we test for compatibility by determining whether $\neg A \wedge \neg B \in \mathcal{D}(\overline{\{A \vee B\}})$. The latter is indeed obtained by Definitions 3 and 9. It follows that $\Sigma, \neg(\{\{A, B\}\} \setminus\!\!\setminus \{A \vee B\}) \mid A \vee B \vdash_{\Theta \cup \overline{A \vee B} \cup \Delta} \Delta$ is defeated and thus that $\Sigma \mid A \vee B \vdash^{\blacktriangleright}_{\Theta} \Delta$ is not a theorem. ■

Remark 2. The consequent of Lemma 9 also holds under the hypothesis that $\Sigma' \mid A, B \vdash_{\Theta'} \Delta$ is a theorem.

Lemma 9 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. $\bigcup \neg(\{\{A \vee B\}\} \setminus\!\!\setminus \{A\}) \in \mathcal{D}(\neg A)$.

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

In order to prevent this unwanted result, we deploy Disjunction Deletion (Definition 11) 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 $\bigcup \neg(\lfloor \mathbf{S}^{\Delta} \rfloor \mathbin{\backslash\!\backslash} \Gamma \cup \{A\}),\, \Sigma \mid \Gamma, A \vdash_{\Theta \cup \Gamma \cup \{A\} \cup \Delta} \Delta$ is the premise in an application of *sturdy*, then $A \vee B \notin \bigcup\{\lfloor \mathbf{S}^{\Delta} \rfloor \mathbin{\backslash\!\backslash} \Gamma \cup \{A\}\}$ for any formula B.

Proof. It suffices to note that according to Definition 11, if $A \in \bigcup \mathbf{S}^{\Delta}$, then $A \vee B \notin \bigcup \lfloor \mathbf{S}^{\Delta} \rfloor$ 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 10, this last constraint enables the set of competitors to be culled. For instance, one can prevent an antecedent, Ω_i, from belonging to $\mathbf{S}^{\Delta}$ by constructing a background set for the candidate that includes information that defeats the inference from Ω_i to Δ. 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 Γ, A is an explanatory argument for Δ, and Δ is the non-trivial consequence of Γ', then together, Γ and Γ' license the inference to A. Since A only forms part of the explanatory argument for Δ, we ought to read *abduct.* as licensing the detachment of a *partial* explanation of Δ. 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 (Δ) 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, it's legitimacy would hardily 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. [JM]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 $\mathsf{LE}^{\blacktriangleright}$ are just the left and right rules for conjunction in LK^{Θ}, *modulo* ►-sequents. We have chosen to restrict the logical operations permitted on ►-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 explanatia/explananda. $^{\mathrm{JM}}$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 ►-sequents. Neither weakening nor cut is permitted. If the antecedents of ►-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 $\Sigma \mid \Gamma \vdash^{\blacktriangleright}_{\Theta} A, \Delta$ and $\Sigma' \mid \Gamma', A \vdash^{\blacktriangleright}_{\Psi} \Delta'$ are theorems of LEA, but $\Sigma', \Sigma \mid \Gamma', \Gamma \vdash_{\Theta\cup\Psi} \Delta, \Delta'$ is not.

Proof. Follows from the fact that there is no cut rule for ►-sequents. ■

While there is no $\mathit{cut}_{\blacktriangleright}$ rule, proofs in LEA contain an application of cut that was not available in LK^{Θ}, 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^{Θ}. The following reduction covers the one application of cut that appears in proofs of LEA that does not appear in proofs of LK^{Θ}.

$$\dfrac{\dfrac{\dfrac{\bigcup\neg(\lfloor\mathbf{S}^{\Delta}\rfloor \setminus\!\!\setminus \Gamma),\ \Sigma \mid \Gamma, A \vdash_{\Theta\cup\overline{\Gamma}\cup\Delta} \Delta}{\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta\cup\overline{\Gamma}\cup\Delta} \Delta}\ \mathit{sturdy} \qquad \begin{matrix}\pi\\ \vdots\\ \Sigma'' \mid \Gamma'' \vdash_{\Psi\cup\overline{\Gamma''}\cup\Delta} \Delta\end{matrix}}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \vdash_{\Theta\cup\overline{\Gamma}\cup\Delta\cup\Psi\cup\overline{\Gamma''}} A}\ \mathit{abduct.} \qquad \Sigma \mid \Gamma, A \vdash_{\Theta} \Delta}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \vdash_{\Theta\cup\overline{\Gamma}\cup\Delta\cup\Psi\cup\overline{\Gamma''}} \Delta}\ \mathit{cut}$$

$$\downarrow$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\begin{matrix}\pi\\ \vdots\\ \Sigma \mid \Gamma'' \vdash_{\Psi} \Delta\end{matrix}}{\Sigma \mid \Gamma'', \Gamma \vdash_{\Psi} \Delta}\,\mathsf{LW}}{\Sigma, A \mid \Gamma'', \Gamma \vdash_{\Psi} \Delta}\,\mathsf{BE}}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \vdash_{\Psi} \Delta}\,\mathsf{BE}}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \vdash_{\Delta\cup\Psi} \Delta}\,\mathsf{DE}}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \vdash_{\overline{\Gamma}\cup\Delta\cup\Psi} \Delta}\,\mathsf{DE}}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \vdash_{\Theta\cup\overline{\Gamma}\cup\Delta\cup\Psi} \Delta}\,\mathsf{DE}}{\Sigma'', \Sigma, A \mid \Gamma'', \Gamma \vdash_{\Theta\cup\overline{\Gamma}\cup\Delta\cup\Psi\cup\overline{\Gamma''}} \Delta}\,\mathsf{DE}$$

■

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. $^{\text{JM}}$Are there broader implications that could drawn out here?

Finally, we can now see that the theorems of LEA constructed with $\vdash^{\blacktriangleright}$ 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.

Proof. (1) Defeasibility follows from Definition 5.

(2) Minimality follows from Lemmas 7, 8, and 9.

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

$$\begin{array}{l} (1)\ \text{If } A \in \mathcal{L} \text{ then } A \in \mathcal{L}_{\blacktriangleright}. \\ (2)\ \text{If } A, B \in \mathcal{L} \text{ then } A \blacktriangleright B \in \mathcal{L}_{\blacktriangleright} \end{array}$$

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^{Θ} apply to all the formulas in $\mathcal{L}_{\blacktriangleright}$, the rules for $\mathsf{LE}^{\blacktriangleright}$ 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.

$$\dfrac{\dfrac{\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} \Delta \qquad \Sigma' \mid \Gamma' \vdash_{\Psi\cup\overline{\Gamma'}\cup\Delta} \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \vdash_{\Theta\cup\Psi\cup\overline{\Gamma'}\cup\Delta} A}\ abduct.}{\Sigma', \Sigma, A \mid \Gamma', \Gamma, A \blacktriangleright B \vdash_{\Theta\cup\Psi\cup\overline{\Gamma'}\cup\Delta} A}\ \mathsf{LW}$$

■

[JM]Is this proposition worth having?

The $\vdash\blacktriangleright$ 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 $\blacktriangleright$-sequent, the conclusion is not. This feature captures the sense in which formulas whose main operator is $\blacktriangleright$ 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 $\blacktriangleright$.

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 LEA^+. In order to do so, we must prove the invertibility of $\vdash\blacktriangleright$. 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

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

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

Rules for ►

$$\dfrac{\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} B, \Delta \qquad \Sigma' \mid \Gamma' \vdash_{\Psi \cup \overline{\Gamma'} \cup \Delta} B, \Delta}{\Sigma', \Sigma, A \mid \Gamma', \Gamma, A \blacktriangleright B \vdash_{\Theta \cup \Psi \cup \overline{\Gamma'} \cup \Delta} A} \blacktriangleright\vdash^{*} \qquad\qquad \dfrac{\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} B, \Delta}{\Sigma, A \mid \Gamma \vdash_{\Theta} A \blacktriangleright B, \Delta} \vdash\blacktriangleright$$

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

Figure 3: Rules of LEA^{+}.

logic (that of the unadorned turnstile, $\vdash_{\Theta}$) encodes the rules of another logic ($\vdash^{\blacktriangleright}_{\Theta}$). To distinguish this claim from the standard deduction theorem, we refer to it as a *quasi-deduction theorem.*

Theorem 3 (Quasi-Deduction Theorem)**.** $\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} B, \Delta$ is provable in LEA^{+} if and only if $\Sigma, A \mid \Gamma \vdash_{\Theta} A \blacktriangleright B, \Delta$ is.

Proof. ($\Rightarrow$) Follows from $\vdash \blacktriangleright$.

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

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

Case 1: If $A \blacktriangleright B$ is not principal in the last rule, it has one or two premises, $\Sigma', [A] \mid \Gamma' \vdash_{\Theta} A \blacktriangleright B, \Delta'$ and $\Sigma'', [A] \mid \Gamma'' \vdash_{\Theta} A \blacktriangleright B, \Delta''$ 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 $\Sigma' \mid \Gamma', A \vdash^{\blacktriangleright}_{\Theta} B, \Delta'$ and $\Sigma'' \mid \Gamma'', A \vdash^{\blacktriangleright}_{\Theta} B, \Delta''$ are obtained by a proof of height n. Now apply the last rule to obtain $\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} B, \Delta$ with a proof of height $\leq n + 1$.

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

JMDiscuss quasi-deduction theorem and show how it captures the spirit of logical expressivism.

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

Theorem 4. Any sequent provable in 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 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.

$$\dfrac{\dfrac{\begin{matrix}\pi_1\\ \vdots\\ \Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} B, \Delta\end{matrix} \qquad \begin{matrix}\pi_2\\ \vdots\\ \Sigma' \mid \Gamma' \vdash_{\Psi\cup\overline{\Gamma'}\cup\Delta} B, \Delta\end{matrix}}{\Sigma', \Sigma, A \mid \Gamma', \Gamma, A \blacktriangleright B \vdash_{\Theta\cup\Psi\cup\overline{\Gamma'}\cup\Delta} A}\;\blacktriangleright\vdash \qquad \dfrac{\Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} B, \Delta}{\Sigma, A \mid \Gamma \vdash_{\Theta} A \blacktriangleright B, \Delta}\;\vdash\blacktriangleright}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \vdash_{\Theta\cup\Psi\cup\overline{\Gamma'}\cup\Delta} A, \Delta}\;\mathit{cut}$$

$$\downarrow$$

$$\dfrac{\dfrac{\begin{matrix}\pi_1\\ \vdots\\ \Sigma \mid \Gamma, A \vdash^{\blacktriangleright}_{\Theta} B, \Delta\end{matrix} \qquad \begin{matrix}\pi_2\\ \vdots\\ \Sigma' \mid \Gamma' \vdash_{\Psi\cup\overline{\Gamma'}\cup\Delta} B, \Delta\end{matrix}}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \vdash_{\Theta\cup\Psi\cup\overline{\Gamma'}\cup\Delta} A}\;\mathit{abduct.}}{\Sigma', \Sigma, A \mid \Gamma', \Gamma \vdash_{\Theta\cup\Psi\cup\overline{\Gamma'}\cup\Delta} A, \Delta}\;\mathsf{RW}$$

■

The conservativity of LEA^+ follows immediately from cut's eliminability.

Corollary 4.1. Every theorem in 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 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 Γ, A is the best *exhaustive* immediate explanation of B, then A is the best *exhaustive* immediate explanation of B *given (all of)* Γ. This reading of the rule corresponds nicely to the pragmatics of explanation. We are rarely in search of *complete* explanations. Rather, our inquires 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 (Γ). In fact, our insistence that defeasible sequents be paired with background sets (Σ) 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 expresssivist 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 fro LK^{Θ} 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 montonicity 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 $\mathsf{LE}^{\blacktriangleright}$?

4. How to represent probabilistic inferences? Treat defeater sets as things that would bring the inference below a threshold probability.