Playing with knowledge and belief
Fiutek, V.

Citation for published version (APA):
Chapter 5

Playing for knowledge

Aim  In this chapter our aim is to provide a game semantics for the “justification games” used to define Keith Lehrer’s notion of “defeasible knowledge”. Indeed while the specifics of Lehrer’s system were not formalized in logical terms and that has left philosophers to settle the misunderstandings via argumentation and on-going philosophical debates, we believe that a formalisation of this type of game can lead to useful insights into the different accounts of knowledge within formal epistemology.

Summary  In this chapter we focus on the many definitions of the concept of “knowledge”. We first present the traditional understanding of knowledge as “justified true belief” as well as Edmund Gettier’s counterexamples to this conception [37]. Then we introduce the notion of defeasible knowledge theorized by Keith Lehrer in [46, 47]. Finally we formalise Lehrer’s conception providing a game semantics for “defeasible knowledge”. The main points are:

– we discuss the various definitions of knowledge, in particular the definition of knowledge as “justified true belief”. We introduce the famous Gettier’s counterexamples shattering this conception of knowledge.

– we develop the notion of “defeasible knowledge” in the form that was theorized by Lehrer. We introduce the “ultra-justification game” as an essential ingredient of Lehrer’s informal account of knowledge as “undefeated justified acceptance”.

– we use total justification models to offer a qualitative representation of an agent’s information and justification within the framework of Dynamic Epistemic Logic.

– Finally we propose a “game semantics for defeasible knowledge”, as a formalization of Lehrer’s conception. We apply our formal model to some
examples, discussing the limits of this formalisation and indicating some possible ways to overcome them.

**Background**  In the traditional literature, there is a common understanding that the famous definition of knowledge as “justified true belief” can/should be credited to Plato in *Meno* and *Theaetetus*. Edmund Gettier himself notes in [37] that Plato seems to consider such a “definition [of knowledge] at *Theaetetus* 201, and perhaps accepting one at *Meno* 98”. However Rohit Parikh rightly points out that Socrates presents an objection to this conception of knowledge in the *Theaetetus*. Thus he shows that contrary to common belief, Plato does not endorse this definition of knowledge.

In [37] Gettier exposes two counter-examples to that definition claiming that having a true justified belief about a given proposition is not sufficient for someone’s knowing this proposition. Indeed these counter-examples show that even true justified beliefs, instead of being real knowledge, can just be lucky guesses. Gettier’s counter-examples are widely accepted by epistemologists as proving that the analysis of knowledge must be modified.

One of the strategies followed by epistemologists to solve the so-called Gettier problem is to find a suitable condition to the definition of knowledge such that knowledge is a justified true belief plus “something”. Keith Lehrer provides such a condition in [46, 47]. Indeed he defines knowledge as undefeated justified acceptance (true belief).

In this chapter we use the setting of justification models as described in Chapter 4.
5.1 Is knowledge justified true belief?

It is generally accepted in contemporary Epistemology that the partitional model of knowledge (first proposed by Hintikka [41] in terms of equivalence relations, and later rediscovered by Aumann) does not provide an adequate picture for “knowledge”, as the term is used in day-to-day life or even in empirical science. From the 1960s when Gettier’s counterexamples [37] shattered the traditional understanding of knowledge as “justified true belief”, many distinguished philosophers proposed various concepts of “knowledge” deemed to be closer to the target. In this section, we focus on the notion of “defeasible knowledge” in the form that was theorized by Lehrer [46, 47].

5.1.1 Discussion about the definition of knowledge

Knowledge as justified true belief The most common interpretation of knowledge is that knowledge is a true belief that can be justified. In other words, knowledge is defined as “justified true belief”. Thus an agent \( S \) knows a proposition \( p \) iff:

1. \( p \) is true
2. \( S \) believes \( p \)
3. \( S \) is justified in believing that \( p \).

The first condition is known as the truth condition and is not controversial. The content of knowledge must be true that is, false propositions cannot be known. The second condition is called the belief condition. Knowledge encodes a propositional attitude towards a proposition. Finally the last condition is the justification condition. Knowledge is not only true belief because an agent could know a proposition being lucky. This agent knows this proposition only if he is able to justify his belief. In other words only if he can provide reasons for his belief.

Gettier problem In [37] Gettier provides some counter-examples to this tripartite analysis of knowledge. We present here a so-called Gettier counter-example. Consider an agent Smith who is justified to believe (1.) that is, he has some reasons to believes (1.).

1. Jones owns a Ford.

Indeed he has evidence that Jones owns a Ford since Jones has offered Smith a ride while driving a Ford, showed Smith some papers stating he owns a Ford and told Smith he owns a Ford.
Now consider an agent Brown who is Smith’s friend. Brown took vacations and went abroad. Smith cannot remember where Brown has gone but it could be Boston, Barcelona or Brest-Litovsk.

And consider the following propositions:

(2.) Either Jones owns a Ford, or Brown is in Boston.
(3.) Either Jones owns a Ford, or Brown is in Barcelona.
(4.) Either Jones owns a Ford, or Brown is in Brest-Litovsk.

Each of these propositions is entailed by (1.). Suppose Smith is a perfect logician. Then Smith is completely justified in believing (2.), (3.) and (4.).

Consider now that Jones does not own a Ford, but an Honda. Besides, by a strange coincidence (and remember entirely unknown to Smith), Brown is indeed in Barcelona. Then (2.) is true, Smith believes that (2.) is true (since he believes that Jones owns a Ford) and he is justified in believing that (2.) is true (since he is justified in believing that Jones owns a Ford). However one cannot claim that Smith knows that (2.) is true.

The Gettier problem is posed in terms of a problem in first order logic. The problem is mainly due to the claim that justification is preserved by entailment: if an agent $S$ is justified to believe $P$ and if $P$ entails $Q$, then $S$ would be justified to believe $Q$. Thus Gettier claims that the three conditions are not sufficient to define knowledge.

**A fourth condition** Several philosophers provide answers to the Gettier problem. Most of them add a fourth condition to the conditions of truth, belief and justification\(^1\). We focus here on the solution provided by Lehrer in [46, 47].

### 5.1.2 Lehrer’s solution to the Gettier problem

Lehrer defines knowledge as undefeated justified acceptance: an agent knows $p$ in case she is justified to accept $p$ and her justification cannot be defeated. He considers several types of justifications varying from a subjective to more objective ones. Both make use of the notions of coherence and reasonableness, but the notion of truth only plays a role in the last one.

---

\(^1\)Some philosophers provide another condition (4.), thus Alvin Goldman who states that a belief is justified only if this belief has been caused by the truth of another belief, considers that a justified true belief is knowledge if the agent is able to correctly reconstruct the causal chain. However, some philosophers prefer to use another notion of justification or to use a primitive notion of knowledge to solve Gettier problem instead of adding a further condition.
5.1. Is knowledge justified true belief?

**Personal justification**  The subjective type of justification is called personal justification. An agent is personally justified to accept \( p \) at time \( t \) iff \( p \) is coherent with the agent's "evaluation system" at \( t \) that is, it is more reasonable for the agent to accept \( p \) than any objection against it on the basis of this evaluation system at \( t \).

**Evaluation system**  The evaluation system of an agent is a collection of three things: an acceptance system, a preference system and a reasoning system. Together this captures both the relevant background information that the agent has acquired about the world in her quest for truth as well as the limited reasoning capacity of the agent. As Lehrer puts it, “the evaluation system of a person consists of what the person accepts, what the person prefers concerning acceptance, and how the person reasons concerning acceptance” [47, p.127]. Then the evaluation system tells what it is more reasonable to accept that is, what sources of information can be trusted (senses, memory...). Note that in the first edition of Lehrer's book, the evaluation system was equated with the acceptance system [46]. But as real agents are not considered to be logically omniscient, it is important in Lehrer's account that one should be able to state the agent's reasoning system explicitly, and similarly the agent's preferences (or conditional acceptances) are an important ingredient when analyzing the agent's knowledge.

**Example of personal justification**  Let's consider an example of a personal justification, that is, an example of coherence with an evaluation system. Imagine I see what looks like a vase on the table in my house, but my reason tells me it could also be a jug of water because they sometimes look like the same. So my senses (my eyes) and my reason disagree on the nature of the object standing on my table. I do not own a jug of water and I accept that I do not own any. But someone could have placed one when I left the house. However I prefer to accept that nobody came into my house than somebody did. Thus my evaluation system (composed of my acceptance that I do not own any jug of water and my preference concerning my acceptance that nobody came into my house) tells me that it is more reasonable to trust my eyes and accept that it is a vase than to accept it is a jug of water. Of course I can be wrong (somebody can actually came to put this jug of water on my table) but I am personally justified to accept that I see a vase.

Hence, an agent who is personally justified to accept a proposition \( p \) at \( t \) might be wrong about the truth-value of \( p \) but will be right in claiming that \( p \) fits well, in the sense of coherence, with the other propositions she accepts. The evaluation system is fallible but according to Lehrer, it has to be used to decide what to accept because relevant information is contained in it. The evaluation system providing personal justification for an acceptance can be in error even though the acceptance itself is true. So we cannot say that a person knows every acceptance
to be true that she is personally justified in accepting provided only that it is also true.

Truth-compatible subsystem  In order to make the transition to a more objective type of justification, Lehrer introduces in [46] the notion of “complete” justification but replaces this in [47] with the idea of justification on the basis of a truth-compatible subsystem of an evaluation system. Such a truth-compatible subsystem of an original evaluation system contains only the accepted items that are actually true, it deletes those states of preference in which something false is preferred over something that is true and its reasoning system is restricted to sound reasonings.

Undefeated justification  This notion of a truth-compatible subsystem plays an essential role in Lehrer’s definition of irrefutable or undefeated justification as follows: an agent is justified to accept \( p \) in a way that is undefeated at \( t \) iff she is justified in accepting \( p \) at \( t \) on the basis of what Lehrer calls the ultra-system at \( t \) (consisting of a truth-compatible subsystem of an original evaluation system at \( t \) and the remaining so-called “unmarked” states). Note the difference in the notions of undefeated justification and personal justification by stressing the role played in the former but not the latter by ingredients that are objectively true.

Justification game  To clarify these notions of justified acceptance, Lehrer defines for each notion of justification a corresponding justification game. In a justification game, an agent (called the Claimant) claims that she is justified to accept \( p \) at time \( t \) while an opponent (called a Skeptic or Critic) tries to show that this is in fact not the case.

Personal justification game  In the personal justification game, the Skeptic can object to the claim of the Claimant by using an objection (or so-called “competitor” as it was called in the first edition [46]) \( o \) to \( p \) iff it is more reasonable for the agent to accept that \( p \) on the assumption that \( o \) is false than on the assumption that \( o \) is true on the basis of her evaluation system at \( t \). Then the Claimant has to answer or neutralize the Skeptic to win a round in the game. If she can answer (show that \( o \) does not cohere with her evaluation system, that is, it is more reasonable for her to accept \( p \) than \( o \) on the basis of her evaluation system) or neutralize (show that there is a neutralizing statement \( n \) which together with \( o \) is not an objection against \( p \) and it is as reasonable to accept \( n \) together with \( o \) as it is to accept \( o \) alone on the basis of her evaluation system) all the objections raised by the Skeptic, she wins the game. If he wins the game, she is personally justified to accept \( p \).
5.1. Is knowledge justified true belief?

Ultra-justification game  In the ultra-justification game, the agent’s payoff is “defeasible knowledge”. In this game the opponent (called Ultra-critic) is supposed to be aware of the truth-value of what the Claimant accepts. The objections can be raised in a similar fashion as in the personal justification game, but now such objections can only be met (answered or neutralized) if they happen to refer to truthful pieces of information (only the content of the truth-compatible subsystem and the existence – but not the content – of the unmarked states of the ultra-system can be used). If the Claimant wins the game, she is justified to accept $p$ in a way that is undefeated.

Defeasible knowledge  If we adopt Lehrer’s definition in [47, p.169] of defeasible knowledge and we use the setting of his justification game to give an explication to condition (4.) below, knowledge will in his setting be reducible to undefeated justified acceptance.

5.1.1. Definition. $S$ knows that $p$ if and only if

1. $S$ accepts that $p$,
2. it is true that $p$,
3. $S$ is justified in accepting that $p$, and
4. $S$ is justified in accepting that $p$ in a way that is not defeated by any false statement (that does not depend on any false statement).

Hence if knowledge of $p$ is reduced to undefeated justified acceptance of $p$, we can say that if the Claimant wins every round of the ultra-justification game then the Claimant knows $p$ (in the (in)defeasible sense of knowledge).

No false lemma  We want to clarify the meaning of condition (4.) above. What does it mean “not defeated by any false statement” and “does not depend on any false statement”? How a justification can be defeated by a false statement?

Lehrer insists on the fact that condition (4.) does not imply the simple denial of false statements. Whenever the false statement is the result of a perceptual error or is the premise of some reasoning, condition (4.) does not imply that the justification the agent has to accept $p$ must not contain any false statements (or beliefs). Moreover Lehrer argues against Peter Klein and Risto Hilpinen’s proposal according to which a justification depends on a false statement iff the person holding the justification would not be justified anymore if she knew the false statement to be false. Lehrer provides examples where knowing some statement to be false is misleading.

2We develop such an example in Section 5.2.3.
In fact, Lehrer only requires that the agent has *some* justification that does not depend on any false statement or is not defeated by any false statement in his definition of knowledge. Here the need and the interest of the ultra-justification game is fully revealed. Remember that this type of game involves using the ultra-system of the agent that is, a system retaining only what is true in the agent’s evaluation system, an evaluation system free of error. Indeed the Ultra-critic can ask the Claimant to eliminate acceptances, preferences and reasonings that do not belong to the ultra-system of the Claimant. The ultra-justification game allows to delete all false statements that could be part of the justification of an agent who can then only use true statements. Then the Claimant wins the game if she can answer or neutralize all the objections of the Ultra-critic proving she is justified to accept \( p \) in a way that is undefeated by any false statement.

If an agent has at least one justification that does not depend on false statement, she will win the ultra-justification game because once she will delete all the false statements belonging to her evaluation system, the Ultra-critic will not be able to defeat the remaining statements and so the corresponding justification\(^3\).

### 5.2 An original game semantics for defeasible knowledge

In this section we formalize Lehrer’s concept of “(in)defeasible knowledge” in terms of a game semantics that we design for this purpose. First, we analyse the notions of belief and knowledge we define in Section 4.5.2 and 4.2.6 from the point of view of Lehrer’s theory of knowledge. Next, we prove Lehrer’s notion of (in)defeasible knowledge (see Definition 5.1.1) to be equivalent to the formal concept of defeasible knowledge.

#### 5.2.1 Knowledge, belief and justification

In Definition 4.2.15, we define the notion of justification: a justification for \( Q \) is an argument \( F \) such that all arguments at least as strong as \( F \) support \( Q \) \((\forall F' (F \leq F' \Rightarrow \bigcap F' \subseteq Q))\). In Lehrer’s personal justification game, this means that the argument \( F \) cannot be defeated by a stronger argument since all stronger arguments support \( Q \). This notion of justification introduced in Chapter 4 corresponds exactly to the notion of personal justification of Lehrer.

In Definition 4.5.2, we define the notion of belief in total justification models: an agent believes \( Q \) iff there exists a justification \( F \) for \( Q \) \((\exists F \forall F' \geq F (\bigcap F' \subseteq Q))\). An agent believes \( Q \) iff all the arguments stronger than \( F \) support \( Q \). In Lehrer’s terminology, this means that the agent is personally justified in accepting \( Q \).

\(^3\)We provide an example to illustrate this in Section 5.2.3.
The notion of justified true belief is captured in our formal system as follows: \( p \land \exists^e \text{just} \, p \). We would like to emphasize here the difference between this definition of justified true belief with our definition of (defeasible) knowledge. We define (defeasible) knowledge in Definition 4.5.2: in total justification models, an agent defeasibly knows \( Q \) at \( s \) if there exists a sound justification \( F \) for \( Q \) at \( s \) \((\exists F(s \in F \land \forall F' \subseteq F((\forall F' \subseteq Q))))\). In our definition of defeasible knowledge, it is not just the belief in \( Q \) that has to be sound but above all the evidence the agent has for \( Q \). An agent defeasibly knows \( Q \) iff the evidence she has for \( Q \) is sound and her evidence supports \( Q \). In Lehrer’s ultra-justification game, this means that the argument \( F \) cannot be defeated by a stronger argument since all stronger arguments support \( Q \) nor by soundness. This notion of justification corresponds exactly to the notion of undefeated justification of Lehrer.

### 5.2.2 Ultra-justification game

Our setting will start from a given justification model \( \mathcal{M} \) which fixes the agent’s justifications, irrevocable knowledge, beliefs, conditional beliefs, strong beliefs and defeasible knowledge. We take this to be the basis of the agent’s evaluation system. From now we only focus on the class of justification models with a total pre-order on bodies of evidence. We distinguish between two kinds of justification models inside this class: the general kind of justification models where the evidence sets are not nested but can be mutually inconsistent or only partially overlapping and the AGM kind where all the evidence sets are nested.

#### Assumptions

We equate Lehrer’s notion of “acceptance” with our notion of “belief” and assume our agent to be logically omniscient. Another assumption we make is that our agent holds only consistent beliefs. These simplifying assumptions render the formalization less complicated and prove sufficient to give a first formal analysis of Lehrer’s justification games. However we are aware of the fact that these restrictions will have to be lifted in future work if we want to have a fully accurate formalization of Lehrer’s account including a formal analysis going beyond the setting provided here.

#### 5.2.1. Definition

Given a total justification model\(^4\) \( \mathcal{M}_0 = (S_0, E_0, \preceq_0, \parallel_0, s_0) \) and a claim \( Q \subseteq S_0 \), we define the ultra-justification game \( G(\mathcal{M}_0, Q) \).

The ultra-justification game is a two players game where the players are called Claimant and Ultra-critic. The justification model of the Claimant is supposed to be known by the Ultra-critic. That means the Ultra-critic knows the epistemic and doxastic attitudes of the Claimant as well as his justification such that the

---

\(^4\)In the rest of the chapter we assume that all justification models and plausibility models are total (i.e. connected) models and we will only explicitly mention the word “total” further on in case confusion is possible.
Ultra-critic knows which belief (and conditional belief, strong belief) is false and which justification is unsound.

Every move for the Claimant (or Believer) is bound by the precondition that the information he conveys to his opponent about himself holding certain justifications, beliefs, conditional beliefs or about the strength of his beliefs, has to be truthful. In other words, the Believer cannot make claims that go against the information he accepts in his own evaluation system, even if what he believes might actually be false in reality.

The pre-condition for any move of the Ultra-critic is that all the information he conveys has to be true in the actual world, i.e. the Ultra-critic cannot lie. This is why we use public announcement operators \(!\) in the formalization of the information that the Ultra-critic conveys.

A play (or run) is a sequence of moves of the players where the set of legal moves of each player is defined below.

The game is played on positions that is, on pairs \(P = (\mathcal{M}, F)\) such that \(\mathcal{M} = (S, E, \preceq, |\cdot|, s)\) is a justification model such that \(s = s_0(\in S)\) and \(F \in \mathcal{E}_\mathcal{M}\) is an argument for that model \(\mathcal{M}\) (i.e. \(F\) is a body of evidence in that model \(\mathcal{M}\)).

The initial position is \(P_0 = (\mathcal{M}_0, F_0)\) where \(\mathcal{M}_0\) is the initial total justification model and \(F_0 = \emptyset\).

The game is played in rounds composed of a move made by the Ultra-critic, followed by a move made by the Claimant. We call the moves of the Ultra-critic “challenges” and the moves of the Claimant “defences”.

Consider a position \(P_{n-1} = (\mathcal{M}_{n-1}, F_{n-1})\) where \(n \geq 1\).

1. First, the Ultra-critic makes a move by:

   a. either challenging the current argument \(F_{n-1}\) as unsound, i.e. announcing that \(s_0 \notin \cap F_{n-1}\) which induces an update \(!(-\cap F_{n-1})\) of the current justification model \(\mathcal{M}_{n-1}\) or

   b. challenging the current argument \(F_{n-1}\) as unconvincing (it is not a justification for \(Q\)), by finding an objection, i.e. an argument \(F' \in \mathcal{E}_{\mathcal{M}_{n-1}}\) such that \(F_{n-1} \preceq F'\) and \(\mathcal{M}_{n-1}, s_0 \models -K(-F' \land \neg Q),\) i.e. \(\cap F' \notin Q \cap S_{n-1}\).

After this, the current justification model is updated to a new justification model \(\mathcal{M}_n\) given by:

\[
\begin{align*}
\mathcal{M}_n := & \begin{cases} 
\mathcal{M}_{n-1}(\neg \cap F_{n-1}) & \text{if the Ultra-critic made a move of type } (a). \\
\mathcal{M}_{n-1} & \text{if the Ultra-critic made a move of type } (b).
\end{cases}
\end{align*}
\]
2. Next, the Claimant correspondingly defends himself by:

\begin{itemize}
\item[a'] answering a challenge of type (a.) with a new argument $F_n \in \mathcal{E}_{\mathcal{M}_n}$, ending at a new position $P_n = (\mathcal{M}_n, F_n)$; or
\item[b'] answering a challenge of type (b.) with a new argument $F_n \in \mathcal{E}_{\mathcal{M}_n}$ such that $F' < F_n$, ending at a new position $P_n = (\mathcal{M}_n, F_n)$.
\end{itemize}

If at any round, a player cannot make a move, then he loses and the other player wins. If the Ultra-critic cannot make a move, i.e. he cannot challenge a given argument, this argument is said to be undefeated. If the Claimant cannot make a move, i.e. he cannot defend a given argument, this argument is said to be defeated. So the Claimant wins the ultra-justification game $G(\mathcal{M}_0, Q)$ iff he offered at least one justification for $Q$ that is left undefeated (his original belief in $Q$ is undefeated).

5.2.2. Proposition. Every play ends in finitely many steps with one of the two players winning.

5.2.3. Proof. Every move of the Claimant either shrinks the total justification model $\mathcal{M}$ or goes to a strictly more convincing argument $F_n$. We know that every total justification model is finite. Since $S_0$ is finite, there are only finitely many updates that shrink the justification model. So there exists some number $n$, such that starting from the $n$-th round, the justification model stays the same forever, i.e. $\forall m > n, \mathcal{M}_n := \mathcal{M}_n$.

From round $n$ onwards, the Claimant can only make moves of type $(b')$. So he can only defend himself providing a new argument such that this argument is strictly more convincing than the argument of the Ultra-critic, which means that at each round, the arguments go stronger: $\forall m > n, F_n < \cdots < F_m$ for every argument $F_n \in \mathcal{E}_{\mathcal{M}_n}$. Since $\mathcal{M}_n$ is finite, $\mathcal{E}_{\mathcal{M}_n}$ is also finite and there are only finitely many available arguments. Hence, there is no infinite ascending chain of arguments, i.e. at some round $m > n$, the Claimant last argument $F_m$ is either defeated (the Ultra-critic wins) or is undefeated (i.e. the Ultra-critic cannot challenge it, that is, he cannot make a move) and the Claimant wins. \(\square\)

5.2.4. Corollary. The game is determined: there exits a winning strategy for one of the players.

5.2.5. Theorem. The Claimant defeasibly knows $Q$ iff he has a winning strategy in the ultra-justification game $G(\mathcal{M}_0, Q)$. Else, the Ultra-critic has a winning strategy in the ultra-justification game $G(\mathcal{M}_0, Q)$.

5.2.6. Proof. Assume the Claimant defeasibly knows $Q$. We are given a total justification model $\mathcal{M}_0$ in which $K_D Q$ is true at $s_0$. We have to show that there exists a winning strategy for the Claimant in $G(\mathcal{M}_0, Q)$. 
We denote by \( L_n \) the (finite) set of available moves for the Claimant at round \( n \) at step (2.), i.e., after the Ultra-critic made his \( n \)-th move and the justification model has been updated to \( \mathcal{M}_n \). Since \( K_P Q \) is true at \( s_0 \) and by Proposition 4.5.11, we know that there exists a sound justification \( F \in \mathcal{E}_0 \) for \( Q \) at \( s_0 \). All we need to show is the following:

**Claim:** for every \( n \), if the step (2.) is reached, i.e., the Ultra-critic made his \( n \)-th move and the justification model has been updated to \( \mathcal{M}_n \), then \( L_n \neq \emptyset \). More precisely, we will show that we will always have

\[
F|\mathcal{M}_n \in L_n
\]

The desired conclusion will follow from this claim. Since for every round \( n \), either the Ultra-critic cannot make his move (hence, he loses) or his challenge can be answered by the Claimant (by choosing any \( F|\mathcal{M}_n \in L_n \)). Hence, the Claimant can never lose. So by Corollary 5.2.4, he will win.

**Proof of the claim:** by induction on \( n \). At round \( n = 0 \), we just have \( L_0 := \mathcal{E}_0 = \{ G \in \mathcal{E}_0 \mid G_0 = \emptyset \subseteq G \} \). So we have \( F \in L_0 \) since \( \emptyset \subseteq F \).

At any later round \( n \), if the Ultra-critic cannot move, the Claimant wins and we are done.

Otherwise if the Ultra-critic made a move of type (a.), announcing !\( (\neg \cap F_{n-1}) \) then we know that we have \( \mathcal{M}_n := \mathcal{M}_{n-1} \setminus (\neg \cap F_{n-1}) \). We want to show that \( F|\mathcal{M}_n \in L_n \), i.e., \( F|\mathcal{M}_n \in \mathcal{E}_{\mathcal{M}_n} \). By induction hypothesis, we know that \( F|\mathcal{M}_{n-1} \in L_{n-1} \subseteq \mathcal{E}_{\mathcal{M}_{n-1}} \). We know that \( F|\mathcal{M}_n := (F|\mathcal{M}_{n-1})|\mathcal{M}_n \). By definition of \( \mathcal{E}_{\mathcal{M}_n} \), \( \mathcal{E}_{\mathcal{M}_n} = \{ G|\mathcal{M}_n \mid G \in \mathcal{E}_{\mathcal{M}_{n-1}} \} \). Then \( F|\mathcal{M}_n := (F|\mathcal{M}_{n-1})|\mathcal{M}_n \in \mathcal{E}_{\mathcal{M}_n} \).

If the Ultra-critic made a move of type (b.), he found an argument \( F' \in \mathcal{E}_{n-1} \) such that \( F_{n-1} \cap F' \cap Q \cap S_{n-1} \), i.e., \( \cap F' \notin Q \). In this case, we know that \( \mathcal{M}_n := \mathcal{M}_{n-1} \). We want to show that \( F|\mathcal{M}_n \in L_n \). By induction hypothesis, we know that \( F|\mathcal{M}_{n-1} \in L_{n-1} \). We have to prove that \( F' \prec_{\mathcal{M}_n} F|\mathcal{M}_n \). Suppose \( F|\mathcal{M}_n \preceq_{\mathcal{M}_n} F' \). By definition of \( \preceq_{\mathcal{M}_n} \), we have \( \{ e \in E_0 \mid e \cap S_n \in F|\mathcal{M}_n \} \preceq_{\mathcal{M}_0} \{ e \in E_0 \mid e \cap S_n \in F' \} \). But \( F \) is sound, i.e., \( s \in e \) for all \( e \in F \), and so \( s \in e \cap S_n \neq \emptyset \) for all \( e \in F \). Hence, \( \forall e \in E_0 (e \in F \Rightarrow e \cap S_n \in F|\mathcal{M}_n \) \). So \( F \subseteq \{ e \in E_0 \mid e \cap S_n \in F|\mathcal{M}_n \} \), and hence, \( F \preceq_{\mathcal{M}_0} \{ e \in E_0 \mid e \cap S_n \in F|\mathcal{M}_n \} \preceq_{\mathcal{M}_0} \{ e \in E_0 \mid e \cap S_n \in F' \} \). Since \( F \) is a justification for \( Q \) in \( \mathcal{M}_0 \), we must have \( \cap \{ e \in E_0 \mid e \cap S_n \in F' \} \in Q \). But \( F' \) is an argument in \( \mathcal{M}_{n-1} := \mathcal{M}_n \). Hence, we have \( e \cap S_n = e \) for all \( e \in F' \) and so \( \cap F' = \cap \{ e \in E_0 \mid e \in F' \} = \cap \{ e \in E_0 \mid e \cap S_n \in F' \} \in Q \), which contradicts the fact that \( F' \) was chosen as an objection by the Ultra-critic (with \( \cap F' \notin Q \)).
For the other direction, assume the Claimant does not defeasibly know \( Q \). We are given total justification model \( \mathcal{M}_0 \) in which \( K_D Q \) is false at \( s_0 \). We want to show that the Ultra-critic has a winning strategy (and hence, the Claimant has not a winning strategy). By Proposition 4.5.11, there does not exist a sound justification \( F \in \mathcal{E}_0 \) for \( Q \) at \( s_0 \).

At every round \( n \), the previous argument \( F_{n-1} \) of the Claimant will be defeated by the Ultra-critic. The Ultra-critic will challenge either by showing that \( F_{n-1} \) is unsound at \( s_0 \) or that \( F_{n-1} \) is unconvincing (it is not a justification for \( Q \)), i.e. by providing an objection \( F' \) with \( F_{n-1} \not\leq F' \) such that \( \bigcap F' \not\subset Q \). The Claimant can never win. Hence, by Corollary 5.2.4, he will lose.

We apply our setting to the following examples. In all the applications of our game semantics, we will use refined justification models as defined in Section 4.2.5, which allow us to distinguish between soft arguments (that weakly support a conclusion, given some implicit biases) and "stronger" arguments (that make the biases explicit).

5.2.3 Applications

The first example is inspired by [47].

Example 1: Zebra

Imagine an agent who is dreaming that she is at the Amsterdam Zoo (Artis) looking at a Zebra. The atomic propositions in this example are zebra (there is a zebra), dream (the agent is dreaming), see (the agent sees a zebra) and Zoo (the agent is at the Amsterdam Zoo). We represent the agent’s evidences via the refined justification model \( \mathcal{M}_0 = \langle S_0, E_0, \models_0, \cdot; \cdot, s_0 \rangle \) described in Figure 5.1.

Note that in this example, we do not consider “seeing” as a factual attitude: what the agent sees is not necessarily true. However, we do consider “seeing” as being fully introspective, that is, if an agent sees something, she irrevocably knows that she sees it and if she does not see something, she irrevocably knows she does not see it. So the agent irrevocably knows that she sees a zebra by introspection. Moreover, the agent irrevocably knows that if she sees a zebra, then either she is at the Zoo or she is dreaming (we assume an agent living in Amsterdam, far away from any savannah). In the same way, she irrevocably knows that if she is at the Amsterdam Zoo, then there is a zebra (we assume she already went to the Zoo where there is indeed a zebra).

In accordance with the knowledge of the agent, her epistemic state consists of four worlds \( S_0 = \{s, t, u, v\} \). The valuation of the atomic propositions is given as follows: zebra is true at \( s, u, v \), dream is true at \( s, t, u \), see is true at \( s, t, u, v \) and
Zoo is true at $u, v$. So at the state $s = s_0$, there is a zebra, the agent sees a zebra, the agent is not at the Zoo and the agent is dreaming. At the state $t$, there is not a zebra, the agent sees a zebra, the agent is not at the Zoo and the agent is dreaming. At the state $u$, there is a zebra, the agent sees a zebra, the agent is at the Zoo and the agent is dreaming. At the state $v$, there is a zebra, the agent sees a zebra, the agent is at the Zoo and the agent is not dreaming. Formally we have:

- $\| \text{zebra} \|_0 = \{s, u, v\}$,
- $\| \text{dream} \|_0 = \{s, t, u\}$,
- $\| \text{see} \|_0 = \{s, t, u, v\}$,
- $\| \text{Zoo} \|_0 = \{u, v\}$.

We remind the reader that, in a refined justification model, not all evidence sets represent genuine evidence. Some evidences sets are biases giving the agent’s default beliefs. In the refined justification model $M_0$, there are four evidence sets:

$$E_0 = \{\text{Zoo}, \text{see}, \neg \text{dream}, \text{zebra} \rightarrow \text{Zoo}\}$$

with:

- $\text{Zoo} = \{u, v\} = e_1$,
- $\text{see} = \{s, t, u, v\} = e_2$,
- $\neg \text{dream} = \{v\} = e_3$,
- $\text{zebra} \rightarrow \text{Zoo} = \{t, u, v\} = e_4$. 
The evidence sets \( e_1 \) and \( e_2 \) represent genuine evidence (through not necessarily truthful). The evidence set \( e_2 \) represents the piece of evidence the agent has, based on her perception: her eyes. The evidence set \( e_1 \) represents the piece of evidence the agent has, based on her memory: she remembers coming to the Zoo. The evidence sets \( e_3 \) and \( e_4 \) represent the biases of the agent. By default, the agent assumes that she is not dreaming, she has no evidence to the contrary so she prefers to believe she is awake (as usually people do). She also assumes that if there is a zebra, then she is at the Amsterdam Zoo (since there is no savannah near Amsterdam).

Note that the evidence set \( e_2 \) has the property that \( e_2 = \text{see} = S_0 \).

We have \( E_0 = \{ \emptyset, \{ e_1 \}, \{ e_2 \}, \{ e_3 \}, \{ e_4 \}, \{ e_1, e_2 \}, \{ e_1, e_3 \}, \{ e_1, e_4 \}, \{ e_2, e_3 \}, \{ e_2, e_4 \}, \{ e_3, e_4 \}, \{ e_1, e_2, e_3 \}, \{ e_1, e_2, e_4 \}, \{ e_1, e_3, e_4 \}, \{ e_2, e_3, e_4 \}, \{ e_1, e_2, e_3, e_4 \} \} \).

The pre-order \( \preceq_0 \) on \( E_0 \) is given by inclusion \( \subseteq \), i.e. \( F \preceq_0 G \) iff \( F \subseteq G \) (so in fact we get an evidence model).

**From refined justification model to plausibility model** We can easily turn this refined justification model into a plausibility model:

1. \( s \preceq_{E_0} t \) iff \( E_t \preceq_0 E_s \) iff \( E_t \subseteq E_s \)
2. \( E_s := \{ e_2 \}, E_t := \{ e_2, e_4 \}, E_u := \{ e_1, e_2, e_4 \} \) and \( E_v := \{ e_1, e_2, e_3, e_4 \} \)
3. \( \{ e_1, e_2, e_4 \} \preceq_0 \{ e_1, e_2, e_3, e_4 \} \) so \( E_u \preceq_0 E_v \)
4. \( \{ e_2, e_4 \} \preceq_0 \{ e_1, e_2, e_4 \} \) so \( E_t \preceq_0 E_u \)
5. \( \{ e_2 \} \preceq_0 \{ e_2, e_4 \} \) so \( E_s \preceq_0 E_t \)

So we have \( E_s \preceq_0 E_t \preceq_0 E_u \preceq_0 E_v \) that is, \( v \preceq_{E_0} u \preceq_{E_0} t \preceq_{E_0} s \).

**Plausibility model** We represent the agent’s beliefs and knowledge via the plausibility model described in Figure 5.2 consisting of four possible states \( (s, t, u, v) \) where the double circled state indicates the real world and the arrows represent the plausibility relation on states (we skip the reflexive and transitive arrows).

One can easily see that our justification model is sphere-based.
The informal dialogue  The dialogue starts with our agent claiming to know that she sees a zebra:
Claimant: There is a zebra here.

\[ B(zebra) \]

Ultra-critic: Why do you think so? (Justify!)

\[ \neg K \neg (\neg zebra) \]

Claimant: I believe there is a zebra because I see a zebra.

\[ \{\text{see}\} \in \mathcal{E}_{M_1} \text{ and } B(\text{see}) \text{ and } B^{\text{see}}(\text{zebra}) \]

Ultra-critic: Maybe you are sleeping and dreaming that you see a zebra.
(Your evidence is consistent with the negation of “zebra”! You need to provide further justification!)

\[ \neg K(\text{see} \land \text{dream} \land \neg zebra) \]

Claimant: It is more reasonable for me to accept that there is a zebra because I see the zebra than to accept that I am dreaming a zebra!

\[ \{\text{see}, \neg \text{dream}\} \in \mathcal{E}_{M_2} \text{ and } B(\text{see} \land \neg \text{dream}) \text{ and } B^{\text{see} \land \neg \text{dream}}(\text{zebra}) \]

Ultra-critic: You are dreaming! You are asleep! You are only seeing the zebra in your dreams!

!\text{dream}

Update  The announcement of “dream” is taken as a public announcement, which formally will change the refined justification model and the plausibility model as described respectively in Figures 5.3 and 5.4:

So \( \mathcal{M}_3 := \mathcal{M}_2(\text{dream}) \) where \( E_{\mathcal{M}_3} = \{\text{zoo, see, dream, zebra }\rightarrow \text{zoo}\} \) with:

- \( \text{Zoo} = \{u\} = e_1 \),
- \( \text{see} = \{s, t, u\} = e_2 \),
- \( \text{zebra }\rightarrow \text{Zoo} = \{t, u\} = e_4 \),
- \( \text{dream} = \{s, t, u\} = e_5 \).

Note that after the update, \( \text{dream} \) is true in all the remaining states \( (S_2(\text{dream}) = \text{dream}) \), so \( \text{dream} \) is an evidence set.

We have \( \mathcal{E}_{\mathcal{M}_3} = \{\emptyset, \{e_1\}, \{e_2\}, \{e_4\}, \{e_3\}, \{e_1, e_2\}, \{e_1, e_4\}, \{e_1, e_3\}, \{e_2, e_4\}, \{e_2, e_3\}, \{e_4, e_3\}, \{e_1, e_2, e_4\}, \{e_1, e_2, e_3\}, \{e_1, e_4, e_3\}, \{e_2, e_4, e_3\}, \{e_1, e_2, e_4, e_5\}\}. \)
The dialogue continued  Then the dialogue continues and our agent claims:
Claimant: I still believe there is a zebra here, coincidental with my dreaming of it. I distinctly remember coming to the Zoo. Maybe I just fell asleep at the Zoo? This would also explain why I am seeing a zebra.

\[ \text{Zoo} \in E_{M_3} \text{ and } B(\text{zoo}) \land B^{\text{zoo}\land\text{dream}\land\text{see}}(\text{zebra}) \]

Ultra-critic: You are not at the zoo. You are asleep in your bed, dreaming of zebras.

\[ !\text{¬Zoo} \]

Update  The announcement of “¬Zoo” is taken as a public announcement, which formally will change the refined justification model and the plausibility model as described respectively in Figures 5.5 and 5.6:

So \( M_4 := M_3(\neg \text{Zoo}) \) where \( E_{M_4} = \{\neg \text{zoo, see, dream, zebra} \rightarrow \text{zoo}\} \) with:

- \( \neg \text{Zoo} = \{s, t\} = e_6, \)
- \( \text{see} = \{s, t\} = e_2, \)
- \( \text{dream} = \{s, t\} = e_5, \)
- \( \text{zebra} \rightarrow \text{Zoo} = \{t\} = e_4 \)
Note again that after the update, \( \neg \text{Zoo} \) is true in all the remaining states \( (S_3(\neg \text{Zoo}) = \neg \text{Zoo}) \), so \( \neg \text{Zoo} \) is an evidence set.

We have \( \mathcal{E}_{M_4} = \{ \emptyset, \{e_2\}, \{e_3\}, \{e_4\}, \{e_5\}, \{e_6\}, \{e_2, e_3\}, \{e_2, e_4\}, \{e_5, e_4\}, \{e_6\}, \{e_4, e_6\}, \{e_2, e_5, e_4\}, \{e_2, e_4, e_5\}, \{e_2, e_4, e_6\}, \{e_5, e_4, e_6\}, \{e_2, e_5, e_4, e_6\} \} \).

**The dialogue ended** Then the dialogue ends with our agent claiming:

Claimant: Given that I am asleep in my bed, I have no justification left to believe there is a zebra here in the bedroom. So I give up: I no longer believe it!

**Conclusion of the dialogue** The agent loses because she cannot provide further argument for \textit{zebra}. She does not even believe \textit{zebra} anymore. She didn’t defeasibly “know” that there was a zebra. In this case our agent does not defeasibly know \textit{zebra} since she based her belief on false evidence. Sadly enough, her initial justified belief was in fact true: however implausible this might seem to her, there is a zebra in her bedroom. The agent did not know: she only had a true justified belief.

**The formal game** Formally, we model the informal dialogue as a play in our ultra-justification game \( G(M_0, zebra) \) as follows.
5.2. An original game semantics for defeasible knowledge

Claimant: There is a zebra here.

\[ B(zebra) \]

This means that at round 0, \( F_0 = \emptyset \).

Ultra-critic: Why do you think so? (Justify!)

\[ \neg K\neg(\neg zebra) \]

This means that at round 1, the Ultra-critic chooses \( F'_0 = F_0 = \emptyset \) such that \( F'_0 \) does not support \( zebra \): \( \bigcap F'_0 = \bigcap \emptyset = S_0 \notin zebra \) because of the state \( t (t \models \neg zebra) \).

Claimant: I believe there is a zebra because I see a zebra.

\( \{\text{see}\} \in E_{M_1} \) and \( B(\text{see}) \) and \( B^{\text{see}}(\text{zebra}) \)

This means that the Claimant chooses a new argument \( F_1 = \{\text{see}\} \) in \( M_1 := M_0 \), which is a soft argument in the sense of Definition 4.2.14 that weakly supports \( zebra \) conditional on the default \( \neg \text{dream} \).

Ultra-critic: Maybe you are sleeping and dreaming that you see a zebra. (Your evidence is consistent with the negation of “zebra”! You need to provide further justification!)

\[ \neg K\neg(\text{see} \land \text{dream} \land \neg zebra) \]

This means that at round 2, the Ultra-critic chooses \( F'_1 = F_1 = \{\text{see}\} \) because \( F'_1 \) does not support \( zebra \): \( \bigcap F'_1 = \bigcap \{\text{see}\} \notin zebra \) because of the state \( t (t \models \text{see} \land \text{dream} \land \neg zebra) \).

Claimant: It is more reasonable for me to accept that there is a zebra because I see the zebra than to accept that I am dreaming a zebra!

\( \{\text{see}, \neg \text{dream}\} \in E_{M_2} \) and \( B(\text{see} \land \neg \text{dream}) \) and \( B^{\text{see}\land\neg \text{dream}}(\text{zebra}) \)

This means that the Claimant makes explicit his bias \( \neg \text{dream} \) by adding it to his argument, obtaining a new argument \( F_2 = \{\text{see}, \neg \text{dream}\} \) in \( M_2 := M_1 \), which is an argument that supports \( zebra \).

Ultra-critic: You are dreaming! You are asleep! You are only seeing the zebra in your dreams!

\( !\text{dream} \)

This means that at round 3, the Ultra-critic challenges \( F_2 \) as unsound and announces that \( s \notin \bigcap F_2 \) which induces an update \(! (\neg \bigcap F_2) \) of the refined justification model and the plausibility model.
**Update** The announcement of “dream” is taken as a public announcement, which formally will change the refined justification model and the plausibility model as described respectively in Figures 5.3 and 5.4:

So $\mathcal{M}_3 := \mathcal{M}_2[\neg \cap F_2]$ where $E_{\mathcal{M}_3} = \{\text{zoo, see, dream, zebra} \rightarrow \text{zoo}\}$ with:

- $\text{Zoo} = \{u\} = e_1$,
- $\text{see} = \{s, t, u\} = e_2$,
- $\text{zebra} \rightarrow \text{Zoo} = \{t, u\} = e_4$,
- $\text{dream} = \{s, t, u\} = e_5$.

Note that after the update, *dream* is true in all the remaining states ($S_2[\neg \cap F_2] = \text{dream}$), so *dream* is an evidence set.

We have $\mathcal{E}_{\mathcal{M}_3} = \{\emptyset, \{e_1\}, \{e_2\}, \{e_4\}, \{e_5\}, \{e_1, e_2\}, \{e_1, e_4\}, \{e_1, e_5\}, \{e_2, e_4\}, \{e_2, e_5\}, \{e_4, e_5\}, \{e_1, e_2, e_4\}, \{e_1, e_2, e_5\}, \{e_1, e_4, e_5\}, \{e_2, e_4, e_5\}, \{e_1, e_2, e_4, e_5\}\}$.

**The game continued** Then the game continues and our agent claims:

Claimant: I still believe there is a zebra here, coincidental with my dreaming of it. I distinctly remember coming to the Zoo. Maybe I just fell asleep at the Zoo?

This would also explain why I am seeing a zebra.

Ultra-critic: You are not at the zoo. You are asleep in your bed, dreaming of zebras.

This means that at round 4, the Ultra-critic challenges $F_3$ as unsound and announces that $s \notin \cap F_3$ which induces an update $!\neg \cap F_3$ of the refined justification model and the plausibility model.

**Update** The announcement of “$\neg \text{Zoo}$” is taken as a public announcement, which formally will change the refined justification model and the plausibility model as described respectively in Figures 5.5 and 5.6:

So $\mathcal{M}_4 := \mathcal{M}_3[\neg \cap F_3]$ where $E_{\mathcal{M}_4} = \{\neg \text{zoo, see, dream, zebra} \rightarrow \text{zoo}\}$ with:

- $\neg \text{Zoo} = \{s, t\} = e_6$,
- $\text{see} = \{s, t\} = e_2$.
5.2. An original game semantics for defeasible knowledge

- dream = \{s,t\} = e_5,
- zebra → Zoo = \{t\} = e_4

Note again that after the update, ¬Zoo is true in all the remaining states \((S_3 \cap F_3) = ¬Zoo\), so ¬Zoo is an evidence set.

We have \(E_{M_4} = \{\emptyset, \{e_2\}, \{e_5\}, \{e_4\}, \{e_2, e_5\}, \{e_2, e_4\}, \{e_2, e_6\}, \{e_5, e_4\}, \{e_4, e_6\}, \{e_2, e_5, e_4\}, \{e_2, e_5, e_6\}, \{e_2, e_4, e_6\}, \{e_5, e_4, e_6\}, \{e_2, e_5, e_4, e_6\}\).

The game ended Then the game ends with our agent claiming:

Claimant: Given that I am asleep in my bed, I have no justification left to believe there is a zebra here in the bedroom. So I give up: I no longer believe it!

Conclusion of the game The agent loses this round of the ultra-justification game because she cannot provide further argument for zebra. She does not even believe zebra anymore. Then the Claimant loses the game: she didn’t defeasibly “know” that there was a zebra. In this case our agent does not defeasibly know zebra since she based her belief on false evidence. Sadly enough, her initial justified belief was in fact true: however implausible this might seem to her, there is a zebra in her bedroom. The agent did not know: she only had a true justified belief.

Example 2: Ferrari

The second example illustrates the meaning of Lehrer’s condition (4.) in his definition of knowledge and underlines the interest of the ultra-justification game. In particular it shows that Lehrer only requires that the agent has at least one justification that does not depend on any false statement to win the ultra-justification game.

Suppose an agent \(S\) is in a room with Mr. Nogot and Mr. Knewit. Mr. Nogot does not own a Ferrari contrary to Mr. Knewit. However the agent \(S\) is justified in accepting that Mr. Nogot owns a Ferrari because \(S\) saw Mr. Nogot drove a Ferrari and Mr. Nogot showed \(S\) the papers stating he owns a Ferrari. Then suppose someone asks \(S\) if she knows whether anyone in the room owns a Ferrari, \(S\) replies claiming she knows that at least one person in the room owns a Ferrari \(P\). It seems that though \(S\) has a justified true belief that \(P\), she does not know it. However, suppose \(S\) also is justified in accepting that Mr. Knewit owns a Ferrari because \(S\) sold Mr. Knewit her Ferrari. Though part of the justification of \(S\) (Mr. Nogot owns a Ferrari) is a false statement/belief, she also has justification that does not depend on this false statement/belief.
The atomic propositions in this example are $P$ (at least one person in the room owns a Ferrari, i.e. $\text{Nogot} \lor \text{Knewit}$), $\text{reliable}$ (Mr. Nogot is reliable), $\text{buy}$ (Mr. Knewit bought the Ferrari of the agent), $\text{Mr. Nogot}$ (Mr. Nogot owns a Ferrari), $\text{Mr. Knewit}$ (Mr. Knewit owns a Ferrari). We represent the agent’s evidence via the refined justification model $\mathcal{M}_0 = (S_0, E_0, \|_0, s_0)$ described in Figure 5.7.

![Figure 5.7: Initial refined justification model](image)

For simplicity, we assume the agent irrevocably knows that if Mr. Nogot is reliable then Mr. Nogot does own a Ferrari while if Mr. Nogot is not reliable, (i.e. Mr. Nogot was lying about the papers stating he owns a Ferrari), he does not actually own a Ferrari\(^6\). We also assume that the agent irrevocably knows that if Mr. Knewit bought her Ferrari then Mr. Knewit does own a Ferrari (since the agent knows Mr. Knewit really wants to own a Ferrari, not to sold one) while if Mr. Knewit did not buy the Ferrari of the agent, then Mr. Knewit does not own a Ferrari (the agent knows that nobody else could have sold one Ferrari to Mr. Knewit).

In accordance with the knowledge of the agent, her epistemic state consists of four worlds $S_0 = s, t, u, v$. The valuation of the atomic propositions is given as follows: $P$ is true at $t, u, v$, $\text{Nogot}$ is true at $u, v$, $\text{Knewit}$ is true at $v, t$, $\text{reliable}$ is true at $u, v$ and $\text{buy}$ is true at $t, v$. So at the state $s$, Mr. Nogot is not reliable, Mr. Nogot does not own a Ferrari, Mr. Knewit did not buy the Ferrari, Mr. Knewit does not own a Ferrari, nobody in the room owns a Ferrari. At the state $t = s_0$, Mr. Nogot is not reliable, Mr. Nogot does not own a Ferrari, Mr. Knewit bought the Ferrari, Mr. Knewit owns a Ferrari, at least one person in the room owns a Ferrari. At the state $u$, Mr. Nogot is reliable, Mr. Nogot owns a Ferrari, Mr. Knewit did not buy the Ferrari, Mr. Knewit does not own a Ferrari, at least one person in the room owns a Ferrari. At the state $v$, Mr. Nogot is reliable,

\(^5\)We assume that our agent knows that she does not own a Ferrari at that moment.

\(^6\)It is not unusual to stop trusting and believing in people when one realize they are liars while one continue to trust them as long as one has evidence that they are telling the truth.
Mr. Nogot owns a Ferrari, Mr. Knewit bought the Ferrari, Mr. Knewit owns a Ferrari, at least one person in the room owns a Ferrari. Formally we have:

- \( P \upharpoonright_0 = \{ t, u, v \} \),
- \( \text{Nogot} \upharpoonright_0 = \{ u, v \} \),
- \( \text{Knewit} \upharpoonright_0 = \{ t, v \} \),
- \( \text{reliable} \upharpoonright_0 = \{ u, v \} \),
- \( \text{buy} \upharpoonright_0 = \{ t, v \} \).

In the refined justification model \( \mathcal{M}_0 \), there are two evidence sets:

\[
E_0 = \{ \text{reliable, buy} \}
\]

with:

- \( \text{reliable} = \{ u, v \} = e_1 \),
- \( \text{buy} = \{ t, v \} = e_2 \).

The evidence set \( e_2 \) represents genuine evidence, i.e. the piece of evidence the agent has, based on her memory: she remembers selling a Ferrari to Mr. Knewit. The evidence set \( e_1 \) represents a bias. By default, she assumes Mr. Nogot is reliable since she has no evidence to the contrary (and we assume the agent prefers to trust people are not liars).

We have the family of bodies of evidence \( \mathcal{E}_0 = \{ \varnothing, \{ e_1 \}, \{ e_2 \}, \{ e_1, e_2 \} \} \).

As pre-order \( \preceq_0 \), we take the cardinality order, i.e. \( F \preceq_0 G \) iff \( |F| \leq |G| \). So the refined justification model \( \mathcal{M}_0 \) is a counting model.

**From refined justification model to plausibility model** We can easily turn this refined justification model into a plausibility model:

1. \( s \preceq_{E_0} t \) iff \( E_t \preceq_0 E_s \) iff \( |E_t| \leq |E_s| \).
2. \( E_t := \{ e_2 \}, E_u := \{ e_1 \} \) and \( E_v := \{ e_1, e_2 \} \)
3. \( \{ e_1 \} \preceq_0 \{ e_1, e_2 \} \) so \( E_u \preceq_0 E_v \)
4. \( \{ e_2 \} \preceq_0 \{ e_1, e_2 \} \) so \( E_t \preceq_0 E_v \)
5. \( \{ e_1 \} \equiv_0 \{ e_2 \} \) so \( E_u \equiv_0 E_t \)

So we have \( E_s \preceq_0 E_u \equiv_0 E_t \preceq_0 E_v \) that is, \( v \preceq_{E_0} t \equiv_{E_0} u \preceq_{E_0} s \).
Plausibility model  We represent the agent’s beliefs and knowledge via the
plausibility model described in Figure 5.8 consisting of four possible states \((s, t, u, v)\)
where the double circled state indicates the real world and the arrows represent
the plausibility relation on states (we skip the reflexive and transitive arrows).

The informal dialogue  The dialogue starts with our agent claiming to know
that someone in the room owns a Ferrari:
Claimant: At least one person in the room owns a Ferrari.

\[ B(P) \]

Ultra-critic: Why do you think so? (Justify!)

\[ \neg K \neg (\neg P) \]

Claimant: I believe at least one person in the room owns a Ferrari because Mr. Nogot is reliable (and he showed me the papers stating he owns a Ferrari).

\[ \{ \text{reliable} \} \in \mathcal{E}_{M_1} \text{ and } B(\text{reliable}) \text{ and } B^{\text{reliable}}(P) \]

Ultra-critic: Mr. Nogot is not reliable, he lied, he does not own a Ferrari!

\[ \neg \neg \text{reliable} \]

Update  The announcement of “\(\neg \text{reliable}\)” is taken as a public announcement,
which formally will change the refined justification model and the plausibility
model as described respectively in Figures 5.9 and 5.10:

So \(M_2 := M_1(\neg \text{reliable})\) where \(E_{M_2} = \{\{e_2\}, \{e_3\}\}\) with:

\[ \begin{aligned}
&\text{buy} = \{t\} = e_2 \\
&\neg \text{reliable} = \{s, t\} = e_3 \\
\end{aligned} \]

Note that after the update, \(\neg \text{reliable}\) is true in all the remaining states
\((S_1)(\neg \text{reliable}) = \neg \text{reliable})\), so \(\neg \text{reliable}\) is an evidence set.

We have \(\mathcal{E}_{M_2} = \{\emptyset, \{e_2\}, \{e_3\}, \{e_2, e_3\}\}\).
5.2. An original game semantics for defeasible knowledge

The dialogue ended

Then the dialogue ends with our agent claiming:

Claimant: I still believe at least one person in the room owns a Ferrari because I remember Mr. Knewit bought mine.

\[ \{ \text{buy} \} \in E_{M_2} \text{ and } B(\text{buy}) \text{ and } B^{\text{buy}}(P) \]

Conclusion of the dialogue

The Ultra-critic cannot object against this last argument because it is actually true that Mr. Knewit bought the Ferrari and so owns one. The Claimant provides a sound justification for \( P \). The Claimant wins, she defeasibly knows at least one person in the room owns a Ferrari because she is justified to accept it in a way that is undefeated by the falsity of any statement (even if her justification for \( P \) contains a false statement/belief).

The formal game

Formally, we model the informal dialogue as a play in our ultra-justification game \( G(M_0, P) \) as follows.

Claimant: At least one person in the room owns a Ferrari.

\[ B(P) \]

This means that at round 0, \( F_0 = \emptyset \).

Ultra-critic: Why do you think so? (Justify!)

\[ \neg K(\neg P) \]
This means that at round 1, the Ultra-critic chooses $F'_0 = F_0 = \emptyset$ such that $F'_0$ does not support $P$: $\cap F'_0 = \cap \emptyset = S_0 \notin P$ because of the state $s (s \equiv \neg P)$.

Claimant: I believe at least one person in the room owns a Ferrari because Mr. Nogot is reliable (and he showed me the papers stating he owns a Ferrari).

\[ \{\text{reliable}\} \in \mathcal{E}_{\mathcal{M}_1} \text{ and } B(\text{reliable}) \text{ and } B^{\text{reliable}}(P) \]

This means that the Claimant chooses a new argument $F_1 = \{\text{reliable}\}$ in $\mathcal{M}_1 := \mathcal{M}_0$, which is an argument that supports $P$.

Ultra-critic: Mr. Nogot is not reliable, he lied, he does not own a Ferrari!

\[ \neg \text{reliable} \]

This means that at round 2, the Ultra-critic challenges $F_1$ as unsound and announces that $t \notin \cap F_1$ which induces an update $!(\neg \cap F_1)$ of the refined justification model and the plausibility model.

**Update** The announcement of “$\neg \text{reliable}$” is taken as a public announcement, which formally will change the refined justification model and the plausibility model as described respectively in Figures 5.9 and 5.10:

So $\mathcal{M}_2 := \mathcal{M}_1 | (\neg \cap F_1)$ where $E_{\mathcal{M}_2} = \{\{e_2\}, \{e_3\}\}$ with:

- $\text{buy} = \{t\} = e_2$
- $\neg \text{reliable} = \{s, t\} = e_3$

Note that after the update, $\neg \text{reliable}$ is true in all the remaining states $(S_1 | (\neg \cap F_1) = \neg \text{reliable})$, so $\neg \text{reliable}$ is an evidence set.

We have $\mathcal{E}_{\mathcal{M}_2} = \{\emptyset, \{e_2\}, \{e_3\}, \{e_2, e_3\}\}$.

**The game ended** Then the game ends with our agent claiming:

Claimant: I still believe at least one person in the room owns a Ferrari because I remember Mr. Knewit bought mine.

\[ \{\text{buy}\} \in \mathcal{E}_{\mathcal{M}_2} \text{ and } B(\text{buy}) \text{ and } B^{\text{buy}}(P) \]

This means that the Claimant chooses a new argument $F_2 = \{\text{buy}\}$ in $\mathcal{M}_2$, which is a sound justification for $P$. 
5.2. An original game semantics for defeasible knowledge

Conclusion of the game  The Ultra-critic cannot object against this last argument because it is actually true that Mr. Knewit bought the Ferrari and so owns one. The Claimant provides a sound justification for $P$. The Claimant wins this run of the game, she defeasibly knows at least one person in the room owns a Ferrari because she is justified to accept it in a way that is undefeated by the falsity of any statement (even if her justification for $P$ contains a false statement/belief).

Example 3: Grabit

The last example is divided into two parts.

First scenario  First, we suppose an agent called Harry who sees a man he knows very well, Tom Grabit, in the library. Harry can see him taking a book and leaving the library without paying.

The atomic propositions in this example are see (Harry saw Tom stealing a book), Tom (Tom stole the book) and Twin (there exists somebody different from Tom who looks just like Tom, that we call “a twin of Tom”). We represent the agent’s evidences via the refined justification model $M_0 = (S_0, E_0, \leq_0, \parallel, s_0)$ described in Figure 5.11.

![Figure 5.11: Initial refined justification model](image)

Note that again, we do not consider “seeing” as a factual attitude but as being fully introspective. So Harry irrevocably knows that he saw “Tom” stealing a book by introspection.

In accordance with the knowledge of the agent, his epistemic state consists of three worlds $S_0 = s, t, u$. The valuation of the atomic propositions is given as follows: see is true at $s, t, u$, Tom is true at $s, u$, Twin is true at $s, t$. So at the state $u = s_0$, Harry saw “Tom” stealing a book and in fact, Tom did steal the book and moreover there does not exist “a twin of Tom”. At the state $t$, Harry saw “Tom” stealing a book, but there exists “a twin of Tom” who actually stole the book (Tom did not steal the book). At the state $s$, Harry saw “Tom” stealing
a book and in fact Tom stole the book, although there exists “a twin of Tom”. Formally we have:
- \( \| see \|_0 = \{s, t, u\} \),
- \( \| Tom \|_0 = \{s, u\} \),
- \( \| Twin \|_0 = \{s, t\} \).

In the refined justification model \( M_0 \), there are two evidence sets:

\[ E_0 = \{\text{see, } \neg Twin\} \]

with:
- \( see = \{s, t, u\} = e_1 \),
- \( \neg Twin = \{u\} = e_2 \).

The evidence set \( e_1 \) represents genuine evidence, i.e. the piece of evidence the agent has, based on his perception: his eyes. The evidence set \( e_2 \) represents a bias of the agent. By default, the agent assumes that there does not exist somebody different from Tom who looks just like Tom since he has no evidence to the contrary (he never met such a person).

Note that the evidence set \( e_1 \) has the property that \( e_1 = see = S_0 \).

We have \( \mathcal{E}_0 = \{\emptyset, \{e_1\}, \{e_2\}, \{e_1, e_2\}\} \).

The pre-order \( \preceq_0 \) on \( \mathcal{E}_0 \) is given by inclusion \( \subseteq \), i.e. \( F \preceq_0 G \) iff \( F \subseteq G \) (so in fact we get an evidence model).\(^7\)

**From refined justification model to plausibility model** We can easily turn this refined justification model into a plausibility model:

1. \( s \preceq_{E_0} t \) iff \( E_t \preceq_0 E_s \) iff \( E_t \subseteq E_s \)
2. \( E_s := \{e_1\}, E_t := \{e_1\} \) and \( E_u := \{e_1, e_2\} \)
3. \( \\{e_1\} \preceq_0 \{e_1, e_2\} \) so \( E_s \preceq_0 E_u \) and \( E_t \preceq_0 E_u \).

So we have \( E_s \equiv_0 E_t \preceq_0 E_u \) that is, \( u \equiv_{E_0} t \preceq_{E_0} s \).

**Plausibility model** We represent the agent’s beliefs and knowledge via the plausibility model described in Figure 5.12 consisting of three possible states \( (s, t, u) \) where the double circled state indicates the real world and the arrows represent the plausibility relation on states (we skip the reflexive and transitive arrows).

\(^7\)Note that the cardinality order would lead to the same conclusions about the defeasible knowledge of the agent.
The informal dialogue  The dialogue starts with our agent claiming to know that Tom stole the book.
Claimant: Tom stole the book.
\[ B(Tom) \]
Ultra-critic: Why do you think so? (Justify!)
\[ \neg K \neg (\neg Tom) \]
Claimant: I believe Tom stole the book because I saw somebody looking just like Tom stealing a book.
\[ \{ \text{see} \} \in \mathcal{E}_{M_1} \text{ and } B(\text{see}) \text{ and } B^{\text{see}}(Tom) \]
Ultra-critic: Maybe there exists “a twin of Tom”!
(Your evidence is consistent with the negation of “Tom”! You need to provide further justification!)
\[ \neg K \neg (\text{see} \land \text{Twin} \land \neg Tom) \]
Claimant: It is more reasonable for me to accept that Tom stole the book because I saw somebody looking just like Tom stealing a book than to accept that there exists “a twin of Tom”!
\[ \{ \text{see}, \neg \text{Twin} \} \in \mathcal{E}_{M_2} \text{ and } B(\text{see} \land \neg \text{Twin}) \text{ and } B^{\text{see} \land \neg \text{Twin}}(Tom) \]

Conclusion of the dialogue  The Ultra-critic cannot object against this last argument because it is actually true that there does not exist “a twin of Tom”. The Claimant provides a sound justification for Tom. The Claimant wins, she defeasibly knows Tom stole a book from the library (even despite the remark of the Ultra-critic about the possible existence of “a twin of Tom”).

The formal game  Formally, we model the informal dialogue as a play in our ultra-justification game \( G(M_0, Tom) \) as follows.
Claimant: Tom stole the book.
\[ B(Tom) \]
This means that at round 0, \( F_0 = \emptyset \).
Ultra-critic: Why do you think so? (Justify!)
\[ \neg K \neg (\neg Tom) \]
Chapter 5. Playing for knowledge

This means that at round 1, the Ultra-critic chooses $F'_0 = F_0 = \emptyset$ such that $F'_0$ does not support $Tom$: $\bigcap F'_0 = \bigcap \emptyset = S_0 \notin Tom$ because of the state $t (t \models \neg Tom)$.

Claimant: I believe Tom stole the book because I saw somebody looking just like Tom stealing a book.

\[
\{\text{see}\} \in \mathcal{E}_{M_1} \text{ and } B(\text{see}) \text{ and } B^{\text{see}}(Tom)
\]

This means that the Claimant chooses a new argument $F_1 = \{\text{see}\}$ in $M_1 := M_0$, which is a soft argument in the sense of Definition 4.2.14 that weakly supports $Tom$ conditional on the default $\neg Twin$.

Ultra-critic: Maybe there exists “a twin of Tom”!
(Your evidence is consistent with the negation of “Tom”! You need to provide further justification!)

\[
\neg K \neg (\text{see} \land Twin \land \neg Tom)
\]

This means that at round 2, the Ultra-critic chooses $F'_1 = F_1 = \{\text{see}\}$ because $F'_1$ does not support $Tom$: $\bigcap F'_1 = \bigcap \{\text{see}\} \notin Tom$ because of the state $t (t \models \text{see} \land Twin \land \neg Tom)$.

Claimant: It is more reasonable for me to accept that Tom stole the book because I saw somebody looking just like Tom stealing a book than to accept that there exists “a twin of Tom”!

\[
\{\text{see}, \neg Twin\} \in \mathcal{E}_{M_2} \text{ and } B(\text{see} \land \neg Twin) \text{ and } B^{\text{see} \land \neg Twin}(Tom)
\]

This means that the Claimant chooses a new argument $F_2 = \{\text{see}, \neg Twin\}$ in $M_2 := M_1$, which is a sound argument that supports $Tom$.

Conclusion of the game  The Ultra-critic cannot object against this last argument because it is actually true that there does not exist “a twin of Tom”. The Claimant provides a sound justification for $Tom$. The Claimant wins this run of the game, she defeasibly knows Tom stole a book from the library (even despite the remark of the Ultra-critic about the possible existence of “a twin of Tom”).

Second scenario  Suppose again that the agent Harry sees a man he knows very well, Tom Grabit, in the library. Harry can see him taking a book and leaving the library without paying. Suppose now that there is really a Twin of Tom, even if it is actually Tom who stole the book.

The atomic propositions in this scenario are exactly the same as those in the first scenario. The refined justification model $M_0 = (S_0, E_0, \preceq_0, \parallel_0, s_0)$ representing the agent’s evidences in this second scenario is identical to the refined justification model in the first scenario with one exception: the actual state $s_0 = s$.

We represent the agent’s beliefs and knowledge via the plausibility model induced from the refined justification model $M_0 = (S_0, E_0, \preceq_0, \parallel_0, s_0)$ described in Figure 5.13.
The informal dialogue  Now the dialogue starts and the Claimant claims to know that Tom stole the book.

Claimant: Tom stole the book.

\[ B(Tom) \]

Ultra-critic: Why do you think so? (Justify!)

\[ \neg K \neg(\neg Tom) \]

Claimant: I believe Tom stole the book because I saw somebody looking just like Tom stealing a book.

\[ \{\text{see}\} \in \mathcal{E}_{M_1} \text{ and } B(\text{see}) \text{ and } B^{\text{see}}(Tom) \]

Ultra-critic: Maybe there exists a Twin of Tom!
(Your evidence is consistent with the negation of “Tom”! You need to provide further justification!)

\[ \neg K(\text{see} \land \text{Twin} \land \neg Tom) \]

Claimant: It is more reasonable for me to accept that Tom stole the book because I saw somebody looking just like Tom stealing a book than to accept that there exists “a twin of Tom”!

\[ \{\text{see}, \neg \text{Twin}\} \in \mathcal{E}_{M_2} \text{ and } B(\text{see} \land \neg \text{Twin}) \text{ and } B^{\text{see} \land \neg \text{Twin}}(Tom) \]

Ultra-critic: There exists a Twin of Tom!

\[ !\text{Twin} \]

Updates  The announcement of “Twin” is taken as a public announcement, which formally will change the refined justification model and the plausibility model as described respectively in Figures 5.14 and 5.15:

So \[ M_3 := M_2[(\text{Twin}) \text{ where } E_{M_3} = \{\text{see},\text{Twin}\} \text{ with } \text{see} = \{s,t\} = e_1 \text{ and } \text{Twin} = \{s,t\} = e_3. \]

Note that after the update, Twin is true in all the remaining states \((S_2)(\text{Twin}) = \text{Twin})\), so Twin is an evidence set.

We have \[ \mathcal{E}_{M_3} = \{\emptyset, \{e_1\}, \{e_3\}, \{e_1, e_3\}\}. \]
Chapter 5. Playing for knowledge

Figure 5.14: Refined justification model updated with !Twin

Figure 5.15: Plausibility model updated with !Twin

The dialogue ended Then the dialogue ends with our agent claiming:
Claimant: Given that there exists a Twin of Tom, I have no justification left to believe Tom stole the book. So I give up: I no longer believe it!

Conclusion of the dialogue The agent loses because she cannot provide further argument for Tom. She does not even believe Tom anymore. Then the Claimant didn’t defeasibly “know” that Tom stole the book. In this case our agent does not defeasibly know Tom since she based her belief on false evidence. Sadly enough, her initial justified belief was in fact true: even if Tom has actually a Twin, he did steal the book. The agent did not know: she only had a true justified belief.

The formal game Formally, we model the informal dialogue as a play in our ultra-justification game $G(M_0, Tom)$ as follows.
Claimant: Tom stole the book.

\[ B(Tom) \]

This means that at round 0, $F_0 = \emptyset$.
Ultra-critic: Why do you think so? (Justify!)

\[ \neg K \neg (\neg Tom) \]
This means that at round 1, the Ultra-critic chooses \( F'_0 = F_0 = \emptyset \) such that \( F'_0 \) does not support \( \text{Tom} \): \( \cap F'_0 = \cap \emptyset = S_0 \not\in \text{Tom} \) because of the state \( t (t \models \neg \text{Tom}) \).

Claimant: I believe Tom stole the book because I saw somebody looking just like Tom stealing a book.

\( \{ \text{see} \} \in \mathcal{E}_{M_1} \) and \( B(\text{see}) \) and \( B^{\text{see}}(\text{Tom}) \)

This means that the Claimant chooses a new argument \( F_1 = \{ \text{see} \} \) in \( \mathcal{M}_1 := \mathcal{M}_0 \), which is a soft argument in the sense of Definition 4.2.14 that weakly supports \( \text{Tom} \) conditional on the default \( \neg \text{Twin} \).

Ultra-critic: Maybe there exists a Twin of Tom!

(Your evidence is consistent with the negation of “Tom”! You need to provide further justification!)

\( \neg K\neg (\text{see} \land \text{Twin} \land \neg \text{Tom}) \)

This means that at round 2, the Ultra-critic chooses \( F'_1 = F_1 = \{ \text{see} \} \) because \( F'_1 \) does not support \( \text{Tom} \): \( \cap F'_1 = \cap \{ \text{see} \} \not\in \text{Tom} \) because of the state \( t (t \models \text{see} \land \text{Twin} \land \neg \text{Tom}) \).

Claimant: It is more reasonable for me to accept that Tom stole the book because I saw somebody looking just like Tom stealing a book than to accept that there exists “a twin of Tom”!

\( \{ \text{see}, \neg \text{Twin} \} \in \mathcal{E}_{M_2} \) and \( B(\text{see} \land \neg \text{Twin}) \) and \( B^{\text{see} \land \neg \text{Twin}}(\text{Tom}) \)

This means that the Claimant chooses a new argument \( F_2 = \{ \text{see}, \neg \text{Twin} \} \) in \( \mathcal{M}_2 := \mathcal{M}_1 \), which is an argument that supports \( \text{Tom} \).

Ultra-critic: There exists a Twin of Tom!

\( !\text{Twin} \)

This means that at round 3, the Ultra-critic challenges \( F_2 \) as unsound and announces that \( s \notin \cap F_3 \) which induces an update \( !(\neg \cap F_2) \) of the refined justification model and the plausibility model.

**Update** The announcement of “Twin” is taken as a public announcement, which formally will change the refined justification model and the plausibility model as described respectively in Figures 5.14 and 5.15:

So \( \mathcal{M}_3 := \mathcal{M}_2(\neg \cap F_2) \) where \( E_{\mathcal{M}_3} = \{ \text{see}, \text{Twin} \} \) with \( \text{see} = \{ s, t \} = e_1 \) and \( \text{Twin} = \{ s, t \} = e_3 \).

Note that after the update, \( \text{Twin} \) is true in all the remaining states \( (S_2)(\neg \cap F_2) = \text{Twin} \), so \( \text{Twin} \) is an evidence set.

We have \( \mathcal{E}_{\mathcal{M}_3} = \{ \emptyset, \{ e_1 \}, \{ e_3 \}, \{ e_1, e_3 \} \} \).
Chapter 5. Playing for knowledge

The game ended Then the game ends with our agent claiming:
Claimant: Given that there exists a Twin of Tom, I have no justification left to believe Tom stole the book. So I give up: I no longer believe it!

Conclusion of the game The agent loses this round of the ultra-justification game because she cannot provide further argument for Tom. She does not even believe Tom anymore. Then the Claimant loses the game: she didn’t defeasibly “know” that Tom stole the book. In this case our agent does not defeasibly know Tom since she based her belief on false evidence. Sadly enough, her initial justified belief was in fact true: even if Tom has actually a Twin, he did steal the book. The agent did not know: she only had a true justified belief.

Conclusion

We provided a formalisation of Lehrer’s ultra-justification game allowing to determine if an agent really (defeasibly) knows some given proposition or if she only believes this proposition. We then proved that an agent defeasibly knows a given proposition iff she continues to believe this proposition as long as she receives only true information (pieces of evidence). We provided the rules for our game semantics for defeasible knowledge stating that an agent defeasibly knows a proposition iff she has a winning strategy in the corresponding formal ultra-justification game.

In the next part, we connect Soft Dynamic Epistemic Logic with other settings. We start by investigating the relations between DEL and the belief revision setting of Dynamic Doxastic Logic in Chapter 6. We show that DDL can internalize all the recent DEL developments for belief revision.