

# Self-referentiality of Justified Knowledge

Roman Kuznets

Ph.D. Program in Computer Science  
CUNY Graduate Center  
365 Fifth Avenue, New York, NY 10016, USA  
kuznets@gmail.com

**Abstract.** The language of justification logic makes it possible to define what it means for knowledge/belief described by an epistemic modality to be self-referential. Building on an earlier result that **S4** and its justification counterpart **LP** describe knowledge that is self-referential, we show that the same is true for **K4**, **D4**, and **T** with their justification counterparts, whereas for **K** and **D** self-referentiality can be avoided. Therefore, no single modal axiom from the standard axiomatizations of these logics can be responsible for self-referentiality.

## 1 Introduction

The modality in **GL** corresponds to provability in the formal arithmetic, which is known to be self-referential. But it is not clear how to formulate this property by means of the modal language itself.

By contrast, the language of justification logic (see [3]) provides a natural way to formulate what it means for the modality in a modal logic to be self-referential. Instead of using existential statements  $\Box F$ , read as “there exists a proof of  $F$ ,” justification logics employ an explicit justification construct  $t : F$ , read “term  $t$  serves as a justification for  $F$ .” In this setting, self-referentiality clearly occurs when a term  $t$  proves something about itself:

$$\vdash t : F(t) . \tag{1}$$

Not only are such constructions allowed by the language, but there are also many theorems of this type, notably with  $t = c$  being an atomic justification, a constant.

**Definition 1.** *Let  $F$  be a justification formula. The forgetful projection  $^\circ$  turns it into a modal formula by replacing each occurrence of justification terms in  $F$  by  $\Box$ ,  $(t : G)^\circ = \Box(G^\circ)$ , while leaving all Boolean connectives and sentence letters intact.*

*The forgetful projection of a set  $X$  of justification formulas is a set of modal formulas  $X^\circ = \{F^\circ \mid F \in X\}$ .*

A logic **L** can be viewed as a set of **L**-theorems. Then

**Definition 2.** A modal logic  $ML$  is said to be a forgetful projection of a justification logic  $JL$  if  $JL^\circ = ML$ .

It was shown in [1] that the forgetful projection of the first justification logic,  $LP$ , is exactly  $S4$ , i.e.,  $LP^\circ = S4$ . This statement is typically called the Realization Theorem because this equality essentially states two things:

1. Replacing each justification term in an  $LP$ -theorem by  $\Box$  yields an  $S4$ -theorem.
2. Vice versa, it is possible to *realize* all occurrences of  $\Box$  in an  $S4$ -theorem by justification terms in such a way that the resulting justification formula is valid. This process of restoring terms hidden in  $\Box$ 's is called *Realization*.

For each of the modal logics  $K, D, T, K4, D4, S4, K5, K45, KD45, S5$  a justification counterpart was developed, so that its forgetful projection is exactly this modal logic (see [1, 3, 4, 8, 9]).

In particular, since  $\vdash \Box A$  for any axiom  $A$  in a modal logic  $ML$ , there must be some term  $t$  in its justification counterpart  $JL$  such that  $\vdash t:(A^r)$ , where  $A^r$  is a realization of  $A$ . In most cases, an axiom of  $ML$  is realized by an axiom of  $JL$ . Justifications for axioms are called *justification constants*, and, unless we have a reason to track or restrict their use, we typically postulate that each constant justifies all axioms. Thus,  $\vdash c:A(c)$ , where  $A(c)$  is an axiom that contains at least one occurrence of  $c$ .

A natural question to ask is **whether such self-referential constants are necessary for the Realization Theorem to hold**. Apart from being direct as in  $\vdash c:A(c)$ , self-referentiality may also occur as a result of a cycle of references:

$$\vdash c_2:A_1(c_1), \quad \dots, \quad \vdash c_n:A_{n-1}(c_{n-1}), \quad \vdash c_1:A_n(c_n) . \quad (2)$$

If direct self-referentiality is expendable, we should ask whether such self-referential cycles are still needed for the Realization.

It was shown in [5] that the realization of  $S4$  in  $LP$  does require direct self-referentiality of constants. In this paper, we prove the following:

- Realization of  $K4$  in  $J4$ , of  $D4$  in  $JD4$ , and of  $T$  in  $JT$  requires direct self-referentiality;
- Realization of  $K$  in  $J$  and of  $D$  in  $JD$  can be performed without any self-referential cycles.

Sect. 2 describes several justification logics and their forgetful projections. At the end of the section, we propose a precise definition of self-referentiality in modal logics. Epistemic semantics for the justification logics from Sect. 2 is described in Sect. 3. Using this semantics, in Sect. 4, we prove that the Realization Theorem for  $K4, D4$ , and  $T$  requires self-referentiality. Sect. 5 demonstrates how to avoid self-referentiality while realizing logics  $K$  and  $D$ . Sect. 6 analyzes the significance of these results and outlines directions for future research.

## 2 Justification Logics and Self-referentiality Defined

The first justification logic, LP, was introduced in [1], where its forgetful projection was shown to be S4 (see also [2]). Justification counterparts for K, D, T, K4, and D4 were developed and the Realization Theorem for them was proven in [4]. Realizations of several modal logics with the Negative Introspection Axiom were considered in [3, 8, 9], but their self-referential properties are outside the scope of this paper, which focuses on the modal logics

$$K, D, T, K4, D4, S4 \quad (3)$$

and their respective justification counterparts

$$J, JD, JT, J4, JD4, LP \quad (4)$$

We will show that for the first two pairs of the modal logic with its justification counterpart self-referentiality can be avoided whereas the last four pairs require direct self-referentiality.

The language of justification logic is that of propositional logic enriched by a new construct  $t:F$ , where  $F$  is any formula and  $t$  is a justification term. Justification terms are built from justification constants  $a, b, c, \dots$  and justification variables  $x, y, z, \dots$  by means of three operations: a unary operation  $!$  and two binary operations  $+$  and  $\cdot$ .<sup>1</sup>

All six justification logics from (4) share the following axioms and rules

- A1. Classical propositional axioms and rule *modus ponens*
- A2. *Application Axiom*  $s:(F \rightarrow G) \rightarrow (t:F \rightarrow (s \cdot t):G)$
- A3. *Monotonicity Axiom*  $s:F \rightarrow (s+t):F, \quad t:F \rightarrow (s+t):F$
- R4. *Axiom Internalization Rule*: for each axiom  $A$  and each justification constant  $c$ , formula  $c:A$  is again an axiom.

These axioms and rules alone yield the basic justification logic J whose forgetful projection is K, the weakest normal modal logic. It is easy to see that the forgetful projection of axioms of J yields theorems of K. Just as other modal logics from (3) are obtained by adding axiom schemes to K, so their justification counterparts from (4) can be obtained by adding corresponding justification schemes to J. In each case, the added modal axiom scheme is the forgetful projection of the respective justification scheme<sup>2</sup>:

Modal Scheme	Justification Scheme	Name of Justification Scheme	To Be Added in Logics
$\Box F \rightarrow F$	$t:F \rightarrow F$	A4. Factivity	JT, LP
$\Box F \rightarrow \Box \Box F$	$t:F \rightarrow !t:(t:F)$	A5. Positive Introspection	J4, JD4, LP
$\Box \perp \rightarrow \perp$	$t:\perp \rightarrow \perp$	A7. Consistency	JD, JD4

<sup>1</sup> Operation  $!$  is only used in J4, JD4, and LP.

<sup>2</sup> Axiom numbering is mostly inherited from [3].

It is important to note that the modal Seriality Axiom in the last row of the table is a single axiom, whereas its realization requires an axiom scheme A7.

**Theorem 1 (Realization Theorem, [1, 4]).**

$$\begin{array}{lll} \text{J}^\circ = \text{K} & \text{JD}^\circ = \text{D} & \text{JT}^\circ = \text{T} \\ \text{J4}^\circ = \text{K4} & \text{JD4}^\circ = \text{D4} & \text{LP}^\circ = \text{S4} \end{array}$$

For each justification logic, a family of weaker logics is defined with a supervised use of rule R4. Note that this rule has a different scope in different justification logics because they have different axiom sets. Thus, the following definition of a constant specification depends on the respective logic. In particular, a constant specification for LP may not be a constant specification for J.

**Definition 3.** A constant specification  $\mathcal{CS}$  for a justification logic  $\mathbf{L}$  is any set of formulas  $c:A$  that can be introduced by the Axiom Internalization Rule R4 of this logic. The only requirement is for such a set to be downward closed, i.e., if  $c_1:c_2:A \in \mathcal{CS}$ , then  $c_2:A \in \mathcal{CS}$ .

**Definition 4.** Let  $\mathcal{CS}$  be a constant specification for a justification logic  $\mathbf{L}$ . By  $\mathbf{L}_{\mathcal{CS}}$  we understand the logic obtained by replacing R4 in logic  $\mathbf{L}$  by the rule

$$\text{R4}_{\mathcal{CS}}. \quad \vdash c:A \quad \text{where } c:A \in \mathcal{CS}.$$

Each logic  $\mathbf{L}$  from (4) is essentially  $\mathbf{L}_{\mathcal{CS}}$  with the total constant specification, i.e., with every constant justifying all axioms.

**Definition 5.** A constant specification  $\mathcal{CS}$  for a justification logic is called self-referential if

$$\{c_2:A_1(c_1), \dots, c_n:A_{n-1}(c_{n-1}), c_1:A_n(c_n)\} \subset \mathcal{CS} \quad (5)$$

for some constants  $c_i$  and axioms  $A_i(c_i)$  with at least one occurrence of  $c_i$ .

A constant specification  $\mathcal{CS}$  is directly self-referential if  $c:A(c) \in \mathcal{CS}$ .

A constant specification is axiomatically appropriate if every axiom  $A$  of the logic has at least one constant  $c$  such that  $c:A \in \mathcal{CS}$ .

The total constant specification is always directly self-referential. Therefore, the standard proofs of the Realization Theorem only show that Realization is possible when direct self-referentiality is used. Our task is to determine whether Realization can be achieved without self-referentiality.

**Definition 6.** Let a modal logic  $\mathbf{ML}$  be the forgetful projection of a justification logic  $\mathbf{JL}$ , i.e.,  $\mathbf{JL}^\circ = \mathbf{ML}$ . We call the modal logic  $\mathbf{ML}$  directly self-referential if  $(\mathbf{JL}_{\mathcal{CS}})^\circ \neq \mathbf{ML}$  for any  $\mathcal{CS}$  that is not directly self-referential.

We call  $\mathbf{ML}$  self-referential if  $(\mathbf{JL}_{\mathcal{CS}})^\circ \neq \mathbf{ML}$  for any  $\mathcal{CS}$  that is not self-referential.

S4 was shown in [5] to be directly self-referential.<sup>3</sup> In this paper, we will prove that K4, D4, and T are also directly self-referential whereas K and D are not self-referential.

<sup>3</sup> The term “directly” was not used in [5].

### 3 Epistemic Models for Justification Logics

Self-referentiality of K4, D4, and T will be established by a semantic argument. Unlike [5], where M-models were used, here we will employ more general F-models, which are based on Kripke models and thus are closer to the standard epistemic semantics. These F-models were first developed for LP; soundness and completeness of LP w.r.t. them can be found in [6]. The adaptation of these models to J, JT, and J4 first appeared in [6]. Soundness and completeness arguments for J and JD can be found in [8], for JT and J4 in [3]. The F-models for JD4 are, perhaps, first developed in this paper.

**Definition 7 (F-models for  $J_{CS}$ ).** An F-model for  $J_{CS}$  is a quadruple  $\mathcal{M} = \langle W, R, \mathcal{A}, v \rangle$ , where  $W \neq \emptyset$  is a set of worlds;  $R \subseteq W \times W$  is an accessibility relation; valuation  $v : SLet \rightarrow 2^W$  assigns to a sentence letter  $P$  a set  $v(P) \subseteq W$  of all worlds where this sentence letter is deemed true; finally, the admissible evidence function  $\mathcal{A} : Tm \times Fm \rightarrow 2^W$  assigns to a pair of a term  $t$  and a formula  $F$  a set  $\mathcal{A}(t, F) \subseteq W$  of all worlds where  $t$  is deemed admissible evidence for  $F$ . The admissible evidence function  $\mathcal{A}$  must satisfy several closure conditions:

- C2.  $\mathcal{A}(t, F \rightarrow G) \cap \mathcal{A}(s, F) \subseteq \mathcal{A}(t \cdot s, G)$
- C3.  $\mathcal{A}(t, F) \cup \mathcal{A}(s, F) \subseteq \mathcal{A}(t + s, F)$
- CS.  $\mathcal{A}(c, A) = W$  for every  $c : A \in CS$ .

The forcing relation  $\Vdash$  is defined as follows:

- $\mathcal{M}, w \Vdash P$  iff  $w \in v(P)$  where  $P$  is a sentence letter;
- Boolean cases are standard;
- $\mathcal{M}, w \Vdash t : F$  iff 1)  $\mathcal{M}, u \Vdash F$  for all  $wRu$  and 2)  $w \in \mathcal{A}(t, F)$ .

The closure conditions C2 and C3 are required to validate axioms A2 and A3 respectively, which is reflected in the numbering. Note that  $w \in \mathcal{A}(t, F)$  in no way implies that  $F$  itself is true. Rather  $w \in \mathcal{A}(t, F)$  means that at world  $w$  term  $t$  is acceptable, although not necessarily conclusive, evidence for  $F$ .

**Definition 8 (F-models for  $JD_{CS}, JT_{CS}, J4_{CS}, JD4_{CS}, LP_{CS}$ ).** An F-model for these logics must satisfy all conditions for an F-model for  $J_{CS}$  plus additional requirements that depend on the additional axioms of the respective logic:

- For  $JT_{CS}$  and  $LP_{CS}$ , axiom  $t : F \rightarrow F$  requires  $R$  to be **reflexive**.
- For  $JD_{CS}$  and  $JD4_{CS}$ , axiom  $t : \perp \rightarrow \perp$  requires  $R$  to be **serial**.
- For  $J4_{CS}, JD4_{CS}$ , and  $LP_{CS}$ , axiom  $t : F \rightarrow !t : t : F$  requires  $R$  to be **transitive**. In addition, two more closure conditions are imposed on  $\mathcal{A}$ :

$$\text{C5.} \quad \mathcal{A}(t, F) \subseteq \mathcal{A}(!t, t : F)$$

**Monotonicity.**  $wRu$  and  $w \in \mathcal{A}(t, F)$  imply  $u \in \mathcal{A}(t, F)$

**Theorem 2 (Completeness Theorem, [3, 6], RK).**  $J_{CS}, JT_{CS}, J4_{CS}$ , and  $LP_{CS}$  are sound and complete w.r.t. their F-models.  $JD_{CS}$  and  $JD4_{CS}$  are sound w.r.t. their F-models; completeness also holds provided CS is axiomatically appropriate.

*Proof.* The cases of  $J_{CS}, JT_{CS}, J4_{CS}$ , and  $LP_{CS}$  are covered in [3]. The proof for  $JD_{CS}$  and  $JD4_{CS}$  can be found in [7].  $\square$

## 4 Self-referential Cases: S4, D4, T, and K4

In [5], direct self-referentiality of knowledge encompassed by S4 and LP was proven by constructing an  $LP_{\mathcal{CS}}$ -counter-model for any potential realization of  $S4 \vdash \diamond(P \rightarrow \Box P)$ , or equivalently, of  $S4 \vdash \neg\Box\neg(P \rightarrow \Box P)$ , where  $\mathcal{CS}$  was the maximal constant specification for LP without directly self-referential constants.

We will employ a similar argument for weaker logics using F-models instead of M-models.

**Theorem 3.** *Realization of D4 in JD4 and of T in JT requires direct self-referentiality.*

*Proof.* Note that  $\Phi = \neg\Box\neg(P \rightarrow \Box P)$  is derivable in both D4 and T.<sup>4</sup> Therefore, we can use the same argument, namely show that no potential realization of  $\Phi$  is valid in  $JD4_{\mathcal{CS}}$ - or  $JT_{\mathcal{CS}}$ -models respectively for the respective maximal  $\mathcal{CS}$  without directly self-referential constants. The proof for these two logics is uniform (and can, in fact, be applied to S4/LP too).

Let  $L \in \{JD4, JT\}$  and  $\mathcal{CS}$  be the maximal constant specification for L without directly self-referential constants. For any pair of terms  $t$  and  $t'$  used in place of the two  $\Box$ 's in  $\Phi$ , we will construct an F-model for  $L_{\mathcal{CS}}$  that falsifies  $\neg t: [\neg(P \rightarrow t': P)]$ , thus showing that no realization of  $\Phi$  is  $L_{\mathcal{CS}}$ -valid. (Note that only soundness is used in this argument.)

Given  $t$  and  $t'$ , consider the following F-model for  $L_{\mathcal{CS}}$ :  $\mathcal{M} = \langle W, R, \mathcal{A}, v \rangle$  with the Kripke frame  $\langle W, R \rangle$  that consists of a single reflexive world  $w$ . Such  $R$  is obviously serial, reflexive, and transitive, thus making the frame suitable for JD4, JT, and LP alike. Let  $v(P) = W = \{w\}$ , i.e.,  $\mathcal{M}, w \Vdash P$ . The truth values of other sentence letters are not important.

Since  $w$  is the only world in the model, we can write  $\Vdash F$  instead of  $\mathcal{M}, w \Vdash F$ ;  $\mathcal{A}(s, F)$  instead of  $w \in \mathcal{A}(s, F)$ ;  $\neg\mathcal{A}(s, F)$  instead of  $w \notin \mathcal{A}(s, F)$ .

The admissible evidence function  $\mathcal{A}$  depends on terms  $t$  and  $t'$ . We require  $\mathcal{A}(t, \neg(P \rightarrow t': P))$ . An admissible evidence function for either logic must satisfy closure conditions C2, C3, and  $\mathcal{CS}$ -closure; additionally for  $JD4_{\mathcal{CS}}$  and  $LP_{\mathcal{CS}}$ , Monotonicity and C5 must hold. Monotonicity is trivially satisfied. Let  $\mathcal{A}$  be the minimal admissible evidence function with  $\mathcal{A}(t, \neg(P \rightarrow t': P))$  that satisfies all the necessary closure conditions. Minimality here means that  $\mathcal{A}(s, F)$  only if it can be derived from  $\mathcal{A}(t, \neg(P \rightarrow t': P))$  using the closure conditions for the logic.

It suffices to show  $\neg\mathcal{A}(t', P)$  to falsify  $\neg t: [\neg(P \rightarrow t': P)]$ . Indeed,  $\not\Vdash t': P$  if  $\neg\mathcal{A}(t', P)$ . Given  $\Vdash P$ , it yields  $\Vdash \neg(P \rightarrow t': P)$ . Finally, with this formula true at the only world and with  $\mathcal{A}(t, \neg(P \rightarrow t': P))$ , we will have  $\Vdash t: [\neg(P \rightarrow t': P)]$ .

$\neg\mathcal{A}(t', P)$  follows from the following technical lemma. Let  $\mathcal{A}_0$  be the minimal admissible evidence function for the logic (without the  $\mathcal{A}(t, \neg(P \rightarrow t': P))$  requirement). Clearly,  $\mathcal{A}_0(s, F)$  implies  $\mathcal{A}(s, F)$  as the closure conditions used are the same, with  $\mathcal{A}$  having one additional *ad hoc* requirement.

**Lemma 1.** *For any subterm  $s$  of term  $t'$ :*

<sup>4</sup> The idea to use this formula for these logics is due to Melvin Fitting.

1. If  $\mathcal{A}_0(s, F)$ , then  $\mathsf{L}_{\mathcal{CS}} \vdash F$  and  $F$  does not contain occurrences of  $t'$ .
2. If  $\mathcal{A}(s, F)$ , but  $\neg\mathcal{A}_0(s, F)$ , then  $F$  has at least one occurrence of  $t'$ . Moreover, the only such implication is  $F = \neg(P \rightarrow t' : P)$ .<sup>5</sup>

*Proof (Sketch).* The proof is by induction on the size of  $s$ . Essentially, we show that all the closures due to C2, an analog of *modus ponens*, happen within  $\mathcal{A}_0$ , so that outside of it the closure derivation is, in a sense, “cut-free.”

The fact that  $\mathcal{CS}$  has no directly self-referential constants is used in the proof of Claim 1 of the lemma: whenever  $\mathcal{A}_0(c, A)$ , we have  $c : A \in \mathcal{CS}$ ; thus, neither  $c$  nor term  $t'$ , whose subterm  $c$  is, can occur in the axiom  $A$ .

The full proof can be found in the Appendix.  $\square$

It remains to apply Lemma 1 to term  $t'$  itself.  $\mathsf{L}_{\mathcal{CS}} \not\vdash P$ , so by Lemma 1.1,  $\neg\mathcal{A}_0(t', P)$ . But then, since  $t'$  does not occur in  $P$ , by Lemma 1.2,  $\neg\mathcal{A}(t', P)$ .  $\square$

**Theorem 4.** *Realization of K4 in J4 requires direct self-referentiality.*

*Proof.* The Hilbert formulation of D4 is obtained from that of K4 by adding the Seriality Axiom. Therefore,  $\mathsf{K4} \vdash \diamond T \rightarrow \diamond(P \rightarrow \Box P)$ ,<sup>6</sup> or equivalently,  $\Psi = \Box\neg(P \rightarrow \Box P) \rightarrow \Box\perp$  is derivable in K4.

For any potential realization  $\Psi^r = t : [\neg(P \rightarrow t' : P)] \rightarrow k : \perp$ , we construct an F-model for  $\mathsf{J4}_{\mathcal{CS}}$  that falsifies  $\Psi^r$ , thus showing that no realization of  $\Psi$  is  $\mathsf{J4}_{\mathcal{CS}}$ -valid. Like in the cases of  $\mathsf{JD4}_{\mathcal{CS}}$  and  $\mathsf{JT}_{\mathcal{CS}}$  from Theorem 3, here  $\mathcal{CS}$  is the maximal constant specification for J4 without directly self-referential constants.

By contrast, the falsifying model here consists of a single irreflexive world. As in such a model any  $F$  is vacuously true at all accessible worlds,  $\Vdash s : F$  iff  $\mathcal{A}(s, F)$ . Again,  $\mathcal{A}$  is taken to be the minimal one with  $\mathcal{A}(t, \neg(P \rightarrow t' : P))$ . Valuation  $v$  is unimportant. We need to show  $\neg\mathcal{A}(k, \perp)$ .

**Lemma 2.** *Let  $\mathcal{A}$  be the minimal admissible evidence function with  $\mathcal{A}(r, B)$  in a single-world F-model for  $\mathsf{J4}_{\mathcal{CS}}$ . If  $\mathcal{A}(s, G)$ , then  $B, r : B \vdash_{\mathsf{J4}_{\mathcal{CS}}} G$ .*

*Proof (Sketch).* The proof is by induction on the closure derivation of  $\mathcal{A}(s, G)$  from  $\mathcal{A}(r, B)$ . It can be easily restored by an interested reader.

The intuition might tell you that  $r : B$  is not necessary as an additional hypothesis. The following example due to Vladimir Krupski shows otherwise: if  $\mathcal{A}(x, P)$  then  $\mathcal{A}(!x, x : P)$ , but surely  $P \not\vdash_{\mathsf{J4}_{\mathcal{CS}}} x : P$ .  $\square$

If  $\mathcal{A}(k, \perp)$ , then, by Lemma 2,  $\neg(P \rightarrow t' : P), t : [\neg(P \rightarrow t' : P)] \vdash_{\mathsf{J4}_{\mathcal{CS}}} \perp$ . But this cannot be the case since in the proof of Theorem 3 we constructed an F-model with both hypotheses being true. It was a  $\mathsf{JD4}_{\mathcal{CS}}$ -model, so it must also be a  $\mathsf{J4}_{\mathcal{CS}}$ -model since fewer restrictions are imposed on the latter and the  $\mathcal{CS}$  for the latter is a subset of the  $\mathcal{CS}$  for the former. A contradiction.  $\square$

<sup>5</sup> We consider  $\neg G$  to be an abbreviation of  $G \rightarrow \perp$ .

<sup>6</sup> The idea to use this formula for K4 is due to Melvin Fitting.

## 5 Non-self-referential Cases: D and K

In this section, we will show that  $(\text{JD}_{\mathcal{CS}})^\circ = \text{D}$  and  $(\text{J}_{\mathcal{CS}})^\circ = \text{K}$  for some non-self-referential constant specifications  $\mathcal{CS}$ .

To construct such constant specifications, we will divide the set of constants into levels indexed by non-negative integers, with each level consisting of countably many constants. Let  $\ell(c)$  denote the level of constant  $c$ . For either logic, let

$$\mathcal{CS} = \{c: A \in \mathcal{TCS} \mid \text{for all constants } a \text{ that occur in } A, \ell(a) < \ell(c)\} . \quad (6)$$

This constant specification is axiomatically appropriate.

**Lemma 3 (Internalization Property).** *Let  $\mathcal{L}_{\mathcal{CS}}$  be a justification logic with an axiomatically appropriate  $\mathcal{CS}$ . Then, for any derivation  $F_1, \dots, F_n \vdash_{\mathcal{L}_{\mathcal{CS}}} B$  there exists an evidence term  $t(x_1, \dots, x_n)$  such that*

$$x_1:F_1, \dots, x_n:F_n \vdash_{\mathcal{L}_{\mathcal{CS}}} t(x_1, \dots, x_n):B . \quad (7)$$

*Proof.* A step-by-step translation from the given derivation into the target one.

$$\begin{array}{ccc} A & \rightsquigarrow & c:A & \text{where } A \text{ is an axiom or } A = c':A' \\ F_i & \rightsquigarrow & x_i:F_i & \text{hypotheses} \\ \frac{D \rightarrow G \quad D}{G} & \rightsquigarrow & \frac{s_1:(D \rightarrow G) \quad s_2:D}{(s_1 \cdot s_2):G} & \text{by A2 and } \textit{modus ponens} \text{ twice} \end{array}$$

□

Since the constant specification (6) has infinitely many constants on each level, it is always possible to choose a fresh constant  $c$  in the second line of the proof.

**Theorem 5.** *It is possible to realize D in JD and K in J without self-referentiality.*

*Proof.* We will prove that  $(\text{JD}_{\mathcal{CS}})^\circ = \text{D}$  and  $(\text{J}_{\mathcal{CS}})^\circ = \text{K}$  for the  $\mathcal{CS}$  from (6). Since  $\mathcal{L}_{\mathcal{CS}} \subseteq \mathcal{L}$ , we have  $(\text{JD}_{\mathcal{CS}})^\circ \subseteq \text{JD}^\circ = \text{D}$  and  $(\text{J}_{\mathcal{CS}})^\circ \subseteq \text{J}^\circ = \text{K}$ .

To show the other inclusion, we will reprove the Realization Theorem using the  $\mathcal{CS}$  from (6). One of the ways to prove Realization is by step-by-step transformation of a cut-free Gentzen derivation of a modal theorem  $F$  into a Hilbert derivation of its realization  $F^r$ . Here  $\vdash \Gamma \Rightarrow \Delta$  is being transformed into  $\Gamma^r \vdash \bigvee \Delta^r$ .<sup>7</sup> A detailed description can be found in [2, 4, 5]. Axioms of the Gentzen modal system are restricted to  $\perp \Rightarrow$  and  $P \Rightarrow P$  for sentence letters  $P$  to have a better control over where and how  $\Box$ 's are introduced. All occurrences of  $\Box$  in the Gentzen modal derivation are divided into families of related occurrences. A cut-free derivation preserves polarity of formulas, so there are positive and negative families of  $\Box$ 's. We realize each negative family by a fresh justification variable. A positive family is realized by a sum of auxiliary variables  $v_1 + \dots + v_n$ , one variable per each use of the modal rules to introduce a  $\Box$  from this family. If all  $\Box$ 's from a positive family are introduced by Weakening, the family is instantiated by a fresh justification variable. The transformation is done by induction on the depth of the Gentzen derivation.

<sup>7</sup> As always, the empty disjunction is interpreted as  $\perp$ .

The Gentzen axioms, propositional rules, and Contraction can be translated using the standard propositional translation from Gentzen into Hilbert. Since the reasoning involved is purely propositional, neither Axiom Internalization is used, nor are new constants introduced. Weakening does not require Axiom Internalization either; it may bring constants from other branches, but never a fresh constant. Thus, new constants are introduced by Axiom Internalization only to translate modal rules. The only modal rule for logic K is  $\frac{C_1, \dots, C_n \Rightarrow B}{\Box C_1, \dots, \Box C_n \Rightarrow \Box B}$ .

In addition, logic D has  $\frac{C_1, \dots, C_n, D \Rightarrow}{\Box C_1, \dots, \Box C_n, \Box D \Rightarrow}$  (see, for instance, [10]). To translate both rules we use the Internalization Property (Lemma 3).

Consider the K-rule first. By IH, we already have a Hilbert derivation of  $C_1^r, \dots, C_n^r \vdash B^r$ . By Lemma 3,  $x_1 : C_1^r, \dots, x_n : C_n^r \vdash t : B^r$  for some  $t$ , where each  $x_i$  is the chosen realization of the negative  $\Box$  in front of  $C_i$ . We then substitute  $t$  for the auxiliary variable that corresponds to this modal rule in the sum realization of the  $\Box$  in front of  $B$  throughout the Hilbert proof.

The D-rule is similar. Here  $x_1 : C_1^r, \dots, x_n : C_n^r, x_{n+1} : D^r \vdash t : \perp$  is obtained after Internalization. Using axiom A7,  $t : \perp \rightarrow \perp$ , and *modus ponens*, we can derive  $\perp$ . Since no positive  $\Box$  is introduced, there is no global substitution of auxiliary variables.

The proof of Lemma 3 shows that the Axiom Internalization Rule in the internalized derivation appears only where axioms or Axiom Internalization Rule instances were in the original derivation. We are free to pick a fresh constant every time. So how can a self-referential cycle appear if we always pick fresh constants? Where does it appear for stronger modal logics? When a term  $t$  substitutes for an auxiliary variable  $v$ , which appears in an Axiom Internalization instance  $c : A(v)$ , the constant  $c$  can *a priori* occur in  $t$ . As shown in Sect. 4 and [5], this cannot be avoided in many logics with other modal Gentzen rules.

We show how to avoid such occurrences of  $c$  in  $t$  for K and D while staying within (6). Let us define the *depth of an occurrence of  $\Box$  in a modal formula  $F$*  by induction on the size of  $F$ : the outer  $\Box$  in  $\Box G$  has depth 0 in  $\Box G$ ; for any occurrence of  $\Box$  inside  $G$ , its depth in  $\Box G$  is obtained by adding 1 to its depth in  $G$ .

Let us also define the *level of an occurrence of  $\Box$  in a Gentzen derivation* as its depth in the formula in which it occurs plus the number of modal rules used on its branch after this occurrence. It is easy to prove that

**Lemma 4.** *In a Gentzen K or D derivation of  $\Rightarrow G$ , levels of all occurrences of  $\Box$  from a given family are equal to the depth of the family's occurrence in  $G$ .*

Let  $N$  be the largest level of  $\Box$ 's in the given cut-free derivation. As we showed, a new constant can be introduced only during Internalization while translating a modal rule. For all rules of level  $i$ , let us always use constants of level  $N - i$ . When constants introduced later on a branch refer to constants introduced on this branch earlier, the former have larger levels because the levels of modal rules decrease toward the root of the derivation. It remains to show that the substitution of terms for auxiliary variables does not violate the level structure of (6).

Indeed, every time a modal rule is used on a branch, all  $\Box$ 's it introduces have the level of this rule, say  $m$ , which is strictly smaller than the levels of all  $\Box$ 's already on the branch. Suppose the Internalization used to translate this modal rule introduced an Axiom Internalization  $c : A(v)$  with an auxiliary variable  $v$ . This  $v$  corresponds to a family of  $\Box$ 's already present on the branch, which must have a larger level  $l > m$ . Wherever the modal rule corresponding to  $v$  occurs, by Lemma 4, it has the same level  $l$ . Therefore, when a term  $t$  substitutes for  $v$ , all the constants in  $t$  will have level  $N - l < N - m = \ell(c)$ . Thus, substitutions do not violate the conditions of our constant specification.  $\square$

## 6 Conclusions and Future Research

Further studies of self-referentiality can develop in various directions. We still do not know an example when self-referentiality is required, but direct self-referentiality can be avoided.

Self-referentiality results can be used to prove structural properties of Gentzen modal derivations, e.g., the unavoidability of double introduction of the same family of  $\Box$ 's on the same branch for directly self-referential modal logics.

It remains to see what triggers self-referentiality. It appears that self-referentiality is tied to the ability to mix levels of  $\Box$ 's in a Gentzen derivation, but we need a larger sample set to make any definite conclusions. We conjecture that the statement of Lemma 4 can be viewed as a purely modal formulation of a sufficient criterion for non-self-referentiality. It would be interesting to see whether it is also necessary.

**Acknowledgements.** The author is greatly indebted to Sergei Artemov, Melvin Fitting, and Vladimir Krupski, whose advice helped to shape this paper. Many thanks to Galina Savukova for editing this text.

## References

- [1] Sergei N. Artemov. Operational modal logic. Technical Report MSI 95–29, Cornell University, 1995.
- [2] Sergei N. Artemov. Explicit provability and constructive semantics. *Bulletin of Symbolic Logic*, 7(1):1–36, 2001.
- [3] Sergei [N.] Artemov. Justification logic. Technical Report TR-2007019, CUNY Ph.D. Program in Computer Science, 2007.
- [4] Vladimir N. Brezhnev. On explicit counterparts of modal logics. Technical Report CFIS 2000–05, Cornell University, 2000.
- [5] Vladimir [N.] Brezhnev and Roman Kuznets. Making knowledge explicit: How hard it is. *Theoretical Computer Science*, 357(1–3):23–34, 2006.
- [6] Melvin Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132(1):1–25, 2005.
- [7] Roman Kuznets. *Complexity Issues in Justification Logic*. PhD thesis, CUNY Graduate Center, 2008.

- [8] Eric Pacuit. A note on some explicit modal logics. In *Proceedings of the 5th Panhellenic Logic Symposium, Athens, Greece, July 25–28, 2005*. University of Athens, 2005.
- [9] Natalia Rubtsova. Evidence reconstruction of epistemic modal logic S5. In *Proceedings of Computer Science Symposium in Russia, CSR 2006*, volume 3967 of *LNCS*, pages 313–321. Springer, 2006.
- [10] Heinrich Wansing. Sequent calculi for normal modal propositional logics. *Journal of Logic and Computation*, 4(2):125–142, 1994.

## Appendix

### Lemma 1

For any subterm  $s$  of term  $t'$ :

1. If  $\mathcal{A}_0(s, F)$ , then  $\mathsf{L}_{\mathcal{CS}} \vdash F$  and  $F$  does not contain occurrences of  $t'$ .
2. If  $\mathcal{A}(s, F)$ , but  $\neg\mathcal{A}_0(s, F)$ , then  $F$  has at least one occurrence of  $t'$ . Moreover, the only such implication is  $F = \neg(P \rightarrow t' : P)$ .

*Proof.* The proof is by induction on the size of  $s$ .

- (A) **Case  $s = \mathbf{x}$** , a justification variable.
    1. For any  $F$ , we have  $\neg\mathcal{A}_0(x, F)$ , so Claim 1 is vacuously true.
    2.  $\mathcal{A}(x, F)$  only if  $t = x$  and  $F = \neg(P \rightarrow t' : P)$ , which does contain  $t'$  and is the only allowed implication.
  - (B) **Case  $s = \mathbf{c}$** , a justification constant.
    1. If  $\mathcal{A}_0(c, F)$ , formula  $F$  must be either an axiom or an instance of the Axiom Internalization Rule. In either case,  $F$  is derivable. At the same time,  $\mathcal{CS}$  is not directly self-referential, so  $F$  cannot contain occurrences of  $c$ , a subterm of  $t'$ . Thus,  $F$  cannot contain  $t'$  either.
    2.  $\mathcal{A}(c, F)$ , but  $\neg\mathcal{A}_0(c, F)$  only if  $t = c$  and  $F = \neg(P \rightarrow t' : P)$ , which does contain  $t'$  and is the only allowed implication.
  - (C) **Case  $s = \mathbf{s}_1 + \mathbf{s}_2$** .
    1. If  $\mathcal{A}_0(s_1 + s_2, F)$ , then, by the closure condition C3,  $\mathcal{A}_0(s_i, F)$  for some  $i = 1, 2$ . By IH,  $F$  is a theorem that does not contain  $t'$ .
    2. If  $\mathcal{A}(s_1 + s_2, F)$ , but  $\neg\mathcal{A}_0(s_1 + s_2, F)$ , then either
      - ( $\alpha$ )  $t = s_1 + s_2$  and  $F = \neg(P \rightarrow t' : P)$ , which satisfies Claim 2, or else
      - ( $\beta$ ) by C3,  $\mathcal{A}(s_i, F)$ , but  $\neg\mathcal{A}_0(s_i, F)$  for some  $i = 1, 2$ . By IH,  $F$  contains  $t'$ , and, if an implication, is  $\neg(P \rightarrow t' : P)$ .
  - (D) **Case  $s = \mathbf{s}_1 \cdot \mathbf{s}_2$** .
    1. If  $\mathcal{A}_0(s_1 \cdot s_2, F)$ , by C2, there must exist a formula  $G$  such that  $\mathcal{A}_0(s_1, G \rightarrow F)$  and  $\mathcal{A}_0(s_2, G)$ . By IH, both  $G \rightarrow F$  and  $G$  are derivable, hence  $F$  is derivable by *modus ponens*. By IH,  $G \rightarrow F$  does not contain  $t'$ , thus neither can  $F$ .
    2. If  $\mathcal{A}(s_1 \cdot s_2, F)$ , but  $\neg\mathcal{A}_0(s_1 \cdot s_2, F)$ , there are several possibilities:
      - ( $\alpha$ )  $t = s_1 \cdot s_2$  and  $F = \neg(P \rightarrow t' : P)$ , which satisfies Claim 2; or else
- by C2, there should exist a  $G$  such that either

( $\beta$ )  $\mathcal{A}(s_1, G \rightarrow F)$  and  $\mathcal{A}(s_2, G)$  while  $\neg\mathcal{A}_0(s_1, G \rightarrow F)$  or

( $\gamma$ )  $\mathcal{A}(s_1, G \rightarrow F)$  and  $\mathcal{A}(s_2, G)$  while  $\neg\mathcal{A}_0(s_2, G)$ .

We will show that both subcases ( $\beta$ ) and ( $\gamma$ ) are inconsistent.

In subcase ( $\beta$ ), by IH, Claim 2 for subterm  $s_1$ ,  $G \rightarrow F = \neg(P \rightarrow t' : P) = (P \rightarrow t' : P) \rightarrow \perp$ . So  $G = P \rightarrow t' : P$ , which is another implication. Hence, by IH, Claim 2 for  $s_2$ , we should have  $\mathcal{A}_0(s_2, G)$ , which contradicts the IH, Claim 1 for  $s_2$  since  $P \rightarrow t' : P$  contains  $t'$ . The contradiction shows impossibility of subcase ( $\beta$ ).

In subcase ( $\gamma$ ), by IH, Claim 2 for  $s_2$ , formula  $G$  should contain  $t'$ . Then  $G \rightarrow F$  would also contain  $t'$ . Hence, by IH, Claim 1 for  $s_1$ , we should have  $\neg\mathcal{A}_0(s_1, G \rightarrow F)$ , and we are back in the impossible subcase ( $\beta$ ). So subcase ( $\gamma$ ) is also impossible.

(E) **Case**  $s = !s_1$  (only for logics  $\mathbf{J4}_{\mathcal{CS}}$ ,  $\mathbf{JD4}_{\mathcal{CS}}$ , and  $\mathbf{LP}_{\mathcal{CS}}$ ).

1. If  $\mathcal{A}_0(!s_1, F)$ , then, by C5,  $F = s_1 : G$  for some  $G$  such that  $\mathcal{A}_0(s_1, G)$ . By IH, Claim 1,  $G$  is a theorem that does not contain  $t'$ .  $\mathcal{A}_0(s_1, G)$  implies that  $\mathcal{A}'(s_1, G) = W$  in any model  $\mathcal{M}' = \langle W', R', \mathcal{A}', v' \rangle$  for  $\mathbf{L}_{\mathcal{CS}}$ . In any such model,  $\mathcal{M}', w' \Vdash G$  for all  $w' \in W'$  by the Soundness part of Theorem 2. By definition of  $\Vdash$ , it follows that  $\mathcal{M}', w' \Vdash s_1 : G$  for any world  $w' \in W'$  in any  $\mathcal{M}'$ . By the Completeness part of Theorem 2,  $s_1 : G$  is derivable.

Since  $G$  does not contain  $t'$  and  $s_1$  is a proper subterm of  $t'$ , formula  $s_1 : G$  cannot contain  $t'$  either.

2. If  $\mathcal{A}(!s_1, F)$ , but  $\neg\mathcal{A}_0(!s_1, F)$ , then either

( $\alpha$ )  $t = !s_1$  and  $F = \neg(P \rightarrow t' : P)$ , which satisfies Claim 2, or else

( $\beta$ ) by C5,  $F = s_1 : G$  for some  $G$  such that  $\mathcal{A}(s_1, G)$ , but  $\neg\mathcal{A}_0(s_1, G)$ . By IH, Claim 2,  $G$  contains  $t'$ , thus so does  $s_1 : G$ , which is not an implication.  $\square$