Counterfactual Reasoning via Grounded Distance

Carlos Aguilera-Ventura , Andreas Herzig , Xinghan Liu , Emiliano Lorini
IRIT-CNRS, University of Toulouse, France
{carlos.aguilera-ventura,andreas.herzig,Emiliano.Lorini}@irit.fr, xinghan.liu@univ-toulouse.fr

Abstract

Conditional logics are usually interpreted in terms of closest world and minimal change. It relies on a measure of distance between worlds which is defined abstractly, i.e. as an element of the model. The typical example of a concrete measure in literature is the Hamming distance. We show that given countably infinite atomic propositions in the language, Hamming distance is not merely an example, but *grounded* for two arguably most important conditional logics, Lewis' VC and VCU. That means, a formula is satisfied in a VC (resp. VCU) model, if and only if it is satisfied in a VC (resp. VCU) model whose distance between worlds is Hammingian.

1 Introduction

Logic of counterfactual conditionals is widely studied and used in different areas including philosophy, linguistics and artificial intelligence. Among many logics of counterfactuals Lewis' VC, VCU and their relatives are arguably the most influential ones.[1] They correspond to, not only other logics of counterfactuals, but also many popular theories in fields e.g. AGM (belief revision) (Grove 1988), KM (belief update) (Grahne 1998), and KLM (preferential reasoning) (Kraus, Lehmann, and Magidor 1990).

One reason for the correspondences is that they can all be subject to some semantics of *minimal change*, the term first used in (Gärdenfors 1984) and later becomes a standard umbrella term for the relative fields, e.g. (Katsuno and Mendelzon 1991; Aiguier et al. 2018). That is to say, for instance, given a counterfactual conditional $\varphi \boxdot\!\!\rightarrow \psi$, if φ is not true at the actual world w, then one should check whether ψ holds at the worlds "closest" to w regarding φ. Closeness in the sense that the change from the actual one must be minimal.

But how is the distance for closeness defined? Most former systems in literature equip with some abstract relation such as epistemic entrenchment, system of spheres, subformulas relations, faithful ordering etc, in order to obtain the additional information to construct the distance measure.

On the other hand, *Hamming distance* is defined as the minimum number of substitutions required to change one string into another. In the context of possible world, it needs no more information than the number of atomic propositions that two worlds disagree with.

Now if we ask for some concrete definition of distance instead of those mysterious ones, Hamming distance is a natural example, e.g. "one candidate of explication . . . is the Hamming distance" (Dizadji-Bahmani and Bradley 2014), "the most commonly used is the Hamming distance" (Aiguier et al. 2018). To our knowledge, this is the much preferred of the only two concrete definitions of distance.[2]

Hence it is not a surprise that a few systems directly define distance as Hammingian. The most famous one is the Dalal operator for belief revision (Dalal 1988), and follow up works are e.g. (Pozos-Parra, Liu, and Perrussel 2013; Delgrande and Peppas 2015). However, a shortcoming is that they all only consider a finite set of atoms/variables. To our limited knowledge, no much literature in AI studies/justifies using Hamming distance given (countably) infinite atoms, though there are some (Williamson 1988; Floridi 2010) in philosophy.

In explainable AI (XAI), the Hamming distance is the "right" distance measure for binary classifier explanation, because the input variables are mutually independent, and counterfactual reasoning is performed by perturbing some variables and observing the output, for studies see e.g. (Darwiche and Hirth 2020; Huang et al. 2022). Recently (Liu and Lorini 2021; Liu and Lorini 2023) present a binary-input classifier logic (BCL) with a conditional operator that essentially corresponds to a version of Lewis' VCU based on the Hamming distance. They show when the language has finite atoms, the conditional operator "is axiomatizable" by reducing to the S5 modal operator. But the case of infinite atoms lacks such property.

Inspired by that undone work, we conjecture *distance being Hamming as a semantic constraint is unaxiomatizable if the language has (countably) infinite atoms*. That indicates that Hamming distance grounds VC and VCU, in the sense that their classes of models satisfy the same set of formulas as their subclasses with Hamming distance.

Section 2 introduces Lewis' V models. In Section 3 we define Hammingian models and have some findings in their own right. The main result is in Section 4, where we show

[1] Through this paper, we use VC to denote both the logic and its axiomatics, VC is the name of the model of VC and **VC** its model class. We do the same for other Lewis' logics mentioned here.

[2] The other one not requiring additional information uses the subset relation: $v \preceq_w u$ iff $V(w)\triangle V(v) \subseteq V(w)\triangle V(u)$, $\triangle$ denoting symmetric difference.

Hamming distance grounds VC and VCU. Section 5 applies the result to BCL for classifier explanation in XAI. Section 6 concludes and discusses.

2 Lewis' V Models

Definition 1. *The language for logics of counterfactual $\mathcal{L}(Atm)$ is defined as follows*[3]

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \boxright \varphi,$$

where p ranges over Atm, a set of countable atomic propositions. Let atm(φ) denote the atoms occurring in φ.

Operators $\vee, \rightarrow, \leftrightarrow$ are defined as usual, and $\bot$ defined as $p \wedge \neg p$, $\top$ as $\neg\bot$, $\Box\varphi$ as $\neg\varphi \boxright \bot$ and $\Diamond\varphi$ as $\neg(\varphi \boxright \bot)$.

We then introduce the comparative similarity model of Lewis' V-logics (V model in short).

Definition 2 (V model)**.** *A tuple $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ is called a V model if W is a non-empty set of worlds, $V : W \longrightarrow 2^{Atm}$ a valuation, and for all $w \in W$, $W_w \subseteq W$ and $\preceq_w$ is a partial order on W_w for comparative similarity with the following constraint:*

- ***Connectedness**: $\forall v, u \in W_w$ either $v \preceq_w u$ or $u \preceq_w v$.*

We have $v \prec_w u$ if $v \preceq_w u$ and $u \npreceq_w v$; $v \approx_w u$ if $v \preceq_w u$ and $u \preceq_w v$. We call M finite if W is finite. The class of V models is noted $\mathbf{V}$.[4]

We note that standard presentations of V models usually do not contain the family of world-indexed sets of accessible worlds W_w but define it from $\preceq_w$ by $W_w =_{def} \{u : \exists v \in W, u \preceq_w v\}$. We prefer to make this component explicit because it will be useful in the rest of the paper.

Definition 3 (Satisfaction relation)**.** *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a V model and $w \in W$:*

$$\begin{aligned}
(M,w) \models p \quad &\Longleftrightarrow \quad p \in V(w);\\
(M,w) \models \neg\varphi \quad &\Longleftrightarrow \quad \textit{it is not } (M,w) \models \varphi;\\
(M,w) \models \varphi \wedge \psi \quad &\Longleftrightarrow \quad (M,w) \models \varphi \textit{ and } (M,w) \models \psi;\\
(M,w) \models \varphi \boxright \psi \quad &\Longleftrightarrow \quad \forall v \in W_w, \textit{ if } (M,v) \models \varphi \textit{ then}\\
&\qquad \exists u \in W_w \textit{ s.t. } 1)\ u \preceq_w v,\\
&\qquad 2)\ (M,u) \models \varphi \wedge \psi, 3)\ \nexists u' \in W_w,\\
&\qquad u' \preceq_w u \textit{ and } (M,u') \models \varphi \wedge \neg\psi.
\end{aligned}$$

Satisfiability and validity are defined in the usual way. We write $\models_{\mathbf{V}} \varphi$ if φ is valid relative to $\mathbf{V}$, that is, if $(M,w) \models \varphi$ for every $M \in \mathbf{V}$ and every w of M.

The satisfaction relation for $\varphi \boxright \psi$, complex as it seems, captures the idea of minimal change by means of $\preceq_w$. Intuitively, $v \preceq_w u$ means that v is closer to w than u; in other words, the distance between v and w is smaller than the distance between u and w. Actually, if the model satisfies the *limit assumption* (defined below), we have a simpler, equivalent satisfaction relation in the light of Lewis' famous equivalence result.[5]

Definition 4 (Selection function)**.** *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a V model, $w \in W$ and $\varphi \in \mathcal{L}(Atm)$. We define $\sigma_w(\varphi)$, the* selection function *of w regarding φ, as*

$$\sigma_w(\varphi) =_{def} \{v \in W_w : (M,v) \models \varphi \ \&\ \forall u \in W_w, \textit{ if } u \neq v \textit{ and } (M,u) \models \varphi, \textit{ then } u \npreceq_w v\}.$$

Definition 5 (Limit assumption)**.** *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a V model. It satisfies the limit assumption, if for all w and φ, if $\exists v \in W_w$ s.t. $(M,v) \models \varphi$ then $\exists u \in W_w$ s.t. $(M,u) \models \varphi$, and $\forall u' \in W_w$ either $u' \npreceq_w u$ or $(M,u') \models \neg\varphi$.*

Fact 1. *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a V model sharing limit assumption and $w \in W$. Then $(M,w) \models \varphi \boxright \psi$ if and only if $\forall v \in \sigma_w(\varphi), (M,v) \models \psi$.*

Since the limit assumption cannot be axiomatized, accepting it or not (Lewis rejected it, unlike most people) actually does not make a substantial difference. However, it echoes in the next section as an interlude.

The V models are too weak to capture most intuitions about counterfactual reasoning. Hence many additional constraints were considered in the literature. The following are welcome, for every w of the given model,

- **Normality (N)**: $W_w \neq \emptyset$;
- **Total reflexivity (T)**: $w \in W_w$;
- **Weak centering (W)**: $w \in W_w$ and $\forall v \in W_w$, $w \preceq_w v$;
- **Centering (C)**: $w \in W_w$ and $\forall v \in W$, if $v \preceq_w w$ then $v = w$;
- **Uniformity (U)**: $W_w = W$.

Most constraints are self-explained. **Weak centering** says that no other world is closer to the current world than itself (but can be equally close); while **Centering** says that the current world is closer to itself than any other world. The hierarchy of the first four is not hard to see, that each one is stronger than the one above it.

From the metaphysical viewpoint, **Centering** is an almost self-evident assumption. Therefore Lewis takes VC, the logic V in adddition with **Centering**, as his "official logic for counterfactuals".[6] **Uniformity** is a desirable additional constraint for VC "in order to forget the bothersome accessibility restrictions and identify the outer modalities with the logical modalities" (Lewis 1995, p. 130). That means, $\Box\varphi$ expresses the universal S5 modality. Lewis names the resulting logic VCU.

By contrast, the following constraints are less desirable:

- **Stalnakerian (S) (Conditional excluded middle):** for each w and φ, either $\sigma_\varphi(w) = \emptyset$ or $|\sigma_\varphi(w)| = 1$;

[3] The reason of the subscript 0 is that we will add a new sort of variables in Section 5.

[4] Unfortunately the *V* for *V* model coincides with the *V* for valuation. While the reader can tell them from context, we mainly discuss about V models with extra properties like VC and VCU, hence mostly no worry of confusion. We do not simply say $\preceq_w$ is a total order to hint that there are weaker models than V models as investigated e.g. in (Burgess 1981).

[5] We ignore the technical issue that σ_w shall take as input the semantic proposition $\|\varphi\|_M =_{def} \{v : (M,v) \models \varphi\}$ instead of the formula φ.

[6] Rigorously speaking, it is V plus the characteristic axiom of **Centering**, similar for other cases. We will see that the model and axiomatic characterizations not always coincide in the next section.

- **Absoluteness (A)**: $\forall w, v \in W, \preceq_w = \preceq_v$.

Lewis calls the first one "Stalnaker's assumption", for [Stalnaker 68] assumes the selection function associates to every world at most one world (and not a set of worlds as the above functions σ_w). It is a bit arbitrary to rule out the possibility that two worlds are equally close to the current one, as illustrated by the famous "if Bizet and Verdi had been compatriots, they would be French" example of (Quine 1950).

As for the second one, it is assumed in some papers in the literature e.g. (Kraus, Lehmann, and Magidor 1990; Goldszmidt and Pearl 1992; Friedman and Halpern 1994) prove that for the complexity of conditional logics "absoluteness makes the problem easier". However, it is such a strong constraint that it becomes unimportant which the indexical/actual world is. Hence (Lewis 1973, p. 131) already says that (to design a logic for counterfactuals) "we surely must reject absoluteness".

Definition 6 (Semantics of subclasses of **V**)**.** *A VX model is a V model satisfying property(ies) X with* $\boldsymbol{X} \subseteq \{\boldsymbol{N, T, W, C, U, S, A}\}$. *The class of VX models is noted* **VX**. *Satisfaction relation, satisfiability and validity in each* **VX** *are defined in the same way as in* **V**.

All the model classes above can be axiomatized in a combinatorial way as the axiomatic of V plus characteristic axioms. But for our main interests we only introduce the axiomatics of VC and VCU.

Definition 7 (Axiomatics of VC and VCU)**.** *The axiomatics of* VCU *is the extension of propositional logic with the following axioms and inference rule. A4 characterizes* ***Weak centering****, A5* ***Centering*** *and A6-7* ***Uniformity****. Hence the axiomatics of* VC *is* VCU *minus A6-7.*

$$\varphi \boxrightarrow \varphi \tag{A1}$$
$$(\varphi \boxrightarrow \neg\varphi) \rightarrow (\psi \boxrightarrow \neg\varphi) \tag{A2}$$
$$((\varphi \boxrightarrow \neg\psi) \vee ((\varphi \wedge \psi) \boxrightarrow \chi)) \leftrightarrow (\varphi \boxrightarrow (\psi \rightarrow \chi)) \tag{A3}$$
$$(\varphi \boxrightarrow \psi) \rightarrow (\varphi \rightarrow \psi) \tag{A4}$$
$$(\varphi \wedge \psi) \rightarrow (\varphi \boxrightarrow \psi) \tag{A5}$$
$$(\varphi \boxrightarrow \bot) \rightarrow (\neg(\varphi \boxrightarrow \bot) \boxrightarrow \bot) \tag{A6}$$
$$\neg(\varphi \boxrightarrow \bot) \rightarrow ((\varphi \boxrightarrow \bot) \boxrightarrow \bot) \tag{A7}$$
$$\frac{(\psi_1 \wedge \cdots \wedge \psi_n) \rightarrow \chi}{((\varphi \boxrightarrow \psi_1) \wedge \ldots (\varphi \boxrightarrow \psi_n)) \rightarrow (\varphi \boxrightarrow \chi)} \tag{RCK}$$

Table 1. Axioms and rule of inference

The last notion to mention is *semantic strength*. Besides comparing two model classes by subset relation, we can say one class is *no weaker than* the other regarding their sets of satisfiable formulas.

Definition 8 (Semantic strength)**.** *Let* $\mathbf{A}, \mathbf{B}$ *be two model classes on the same language. By* $\mathbf{A} \sqsubseteq \mathbf{B}$ *we denote for every* φ*, if* φ *is satisfiable in* $\mathbf{A}$*, then* φ *is satisfiable in* $\mathbf{B}$*; by* $\mathbf{A} \sqsubset \mathbf{B}$ *we denote* $\mathbf{A} \sqsubseteq \mathbf{B}$ *but not* $\mathbf{B} \sqsubseteq \mathbf{A}$*; by* $\mathbf{A} \equiv \mathbf{B}$ *we denote both* $\mathbf{A} \sqsubseteq \mathbf{B}$ *and* $\mathbf{B} \sqsubseteq \mathbf{A}$ *and call them* equivalence*.*

Notice that if $\mathbf{A} \subseteq \mathbf{B}$ then $\mathbf{B} \sqsubseteq \mathbf{A}$, but the inverse does not necessarily hold. In particular, possibly $\mathbf{A} \neq \mathbf{B}$ and $\mathbf{A} \equiv \mathbf{B}$.

3 Hammingian Models for Counterfactuals

3.1 Hammingian Lewis Models

In a certain sense, Lewis' models are Kripke models plus relations of comparative similarity. A natural question is: closeness (similarity) according to what measure? As mentioned in literature, the most concrete and almost standard example in the literature is closeness in sense of the *Hamming distance* between possible worlds.

Definition 9 (Hamming distance between worlds)**.** *Let* W *be a non-empty set of worlds and* $V : W \longrightarrow 2^{Atm}$*. For any* $w, v \in W$*, their Hamming distance under* V *is defined as* $\hbar_V(w, v) = |V(w) \triangle V(v)|$*, where* $\triangle$ *denotes symmetric difference.*

Definition 10 (Hammingian V model)**.** *A V model* $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ *is Hammingian if* $\forall v, u \in W_w$*,* $v \preceq_w u$ *iff* $\hbar_V(w, v) \leq \hbar_V(w, u)$*. The class of HV models is noted* **HV**. *The subclasses of HV models are defined and noted in the similar way as V models.*

Interestingly, the disputation of accepting limit assuption or not does not bother us in **HV**.

Fact 2. *Every HV model satisfies the limit assumption.*

Although the limit assumption is closely related to well-foundedness, it is not the case that $\preceq_w$ is well-founded. Indeed, let $p_1, p_2, \ldots$ be some enumeration of the atoms of Atm and let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be the HVU model where $W = Atm \cup \{p_1, \ldots, p_n : n \in \mathbb{N}\}$ and V is identity. Then $\{p_1\} \succ_{Atm} \{p_1, p_2\} \succ_{Atm} \{p_1, p_2, p_3\} \succ_{Atm} \ldots$ is an infinite descending chain.

A special case are HV models containing all logically possible worlds, i.e., all elements of 2^{Atm}. This corresponds to the semantics of (Dalal 1988) for database updates. For that semantics, Π_2^p completeness of deciding whether $\varphi \rightarrow (\psi \boxrightarrow \chi)$ was proved in (Eiter and Gottlob 1992), and the validities were axiomatized in (Herzig 1998).

3.2 Model (Sub)classes: a Comparison

Subset relations between V model subclasses are shown in (Lewis 1973, Figure 5, p. 131), where semantic strength relations are just inverses of the former. We will see that in **HV**, Hamming distance not only determines the comparative similarity, but also "perturbs" the constraints of V models. Consequently, more relations between subclasses of **HV** can be found. Particularly, subset and semantic strength relations no more just inverse. A summary is in Figure 1.

Proposition 1. **HVT** = **HVW**.

Proof. Inherited from the V models, **HVT** $\supseteq$ **HVW**. For the other direction, let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be an HVT model. Then by the Hamming distance obviously $\forall w \in W, \forall v \in W_w, w \preceq_w v$, i.e. M is weakly centered. □

Similarly, the fact below is easy to see.

Fact 3. **HVU** $\subset$ **HVW**.

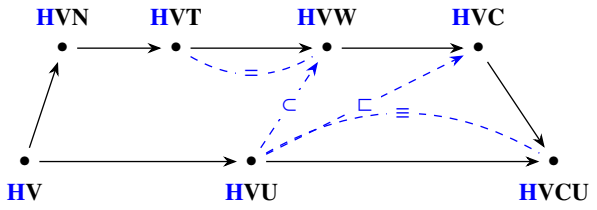

Figure 1. Model class relations. Black parts are results of Lewis where arrow means subset relation between model classes; blue parts are new findings of their Hammingian subclasses, where each dash line means the relation by its indicator in $\{=, \equiv, \subset, \sqsubset\}$.

Fact 4 (Indiscernibility)**.** *Let* $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ *be an HVU model. Then* $\forall w, v \in W$, $V(w) = V(v)$, *if and only if* $v \preceq_w w$ *if and only if* $\preceq_w = \preceq_v$.

Proposition 2. $\mathbf{HVC} \supset \mathbf{HVU}$.

Proof. Suppose $M \in \mathbf{HVU}$. First, observe that for every $w \in W$, the closest worlds around w is all those worlds v such that $v \preceq_w w$. Second, by Fact 4 $v \preceq_w w$ implies that $V(v) = V(w)$ and $\preceq_v = \preceq_w$. Third, we define a relation $Z \subseteq W \times W$ by: wZv iff $V(w) = V(v)$. Thanks to the second observation, it is not hard to see that Z is a bisimulation. It follows that all closest worlds around w satisfy the same formulas. Hence the centering axiom is valid in **HVU** models. □

Noticing that $\mathbf{HVCU} \subset \mathbf{HVU}$, the following result becomes obvious.

Fact 5. $\mathbf{HVU} \equiv \mathbf{HVCU}$.

3.3 Hamming State Models

One can argue conceptually that Hamming distance commits us to identify a world with its valuation, or even stronger, that the real model shall be defined by valuations of variables rather than more abstract entities, namely worlds. Formally we can define the following.

Definition 11 (Hamming state model)**.** *We call a model* $\mathcal{S} = (S, (S_s)_{s\in S})$ a Hamming state model with parameter[7] *if* $S \subseteq 2^{Atm}$ *and* $\forall s \in S, S_s \subseteq S$. *If* $S_s = S$ *for each* $s \in S$, *we just call it a* Hamming state model.

The use of the Hamming distance is justified by Leibniz's law in (Floridi 2010). For any w, v in a model, they are identical if $w = v$; equivalent if $V(w) = V(v)$; indiscernible if $w \approx_u v$ for every u in the model. Fact 4 states the indiscernibility of equivalences in **HVU**. **Centering** states a stronger property: the identity of equivalences in **HVU** (but with a restriction to accessible worlds). Thus we can examine the philosophy from a more logical viewpoint in terms of bisimilarity and isomorphism.

Fact 6. *Let* $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ *be an HVC model. Then* $\forall v \in W_w$, $V(w) = V(v)$ *if and only if* $w = v$. *Particularly, if M is an HVCU model then* $\forall w, v \in W, V(w) = V(v)$ *if and only if* $w = v$.

[7] We call this one "Hamming", while the adjective "Hammingian" qualifies HV models.

Proposition 3. *Every HVU model is bisimilar to a Hamming state model with parameter; every HVCU model is isomorphic to a Hamming state model.*

Proof. Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be an HVU model. For each $w \in W$, let s_w denote $V(w)$. Consider $\mathcal{S} = (S, (S_s)_{s\in S})$ s.t. $S = \{s_w : w \in W\}$ and $\forall s_w \in S, s_v \in S_{s_w}$ iff $v \in W_w$. The bisimulation between the two models is obvious. The same construction applies to HVCU but the result is stronger because of Fact 6. □

The above result for HVCU does not hold for weaker logics: for every $\mathbf{X} \in \{\mathbf{N}, \mathbf{T}, \mathbf{W}, \mathbf{C}\}$ there is a model in **HVX** that is not bisimular to any Hamming state model with parameter. To see that, simply consider an HVX model with two worlds w, v, s.t. $V(w) = V(v)$, $W_w = \{w\}$ and $W_v = \{v\}$.

Despite the isomorphism between HVCU and state models, we keep using possible worlds semantics in line with other HV models until Section 6 when state models debut.

4 Equivalence Results Given Infinite Atoms

Now that the stage is set, let us raise the bold question: is the Hamming distance not just an example, but *the* grounded measure of distance for VC and VCU? Grounded in the intuitive sense that, given an infinite supply of atoms, we can transform any non-Hammingian model to a Hammingian one while preserving the truth of some formula. Formally the question is put as the following theses of equivalences.

Thesis 1. $\mathbf{VC} \equiv \mathbf{HVC}$.

Thesis 2. $\mathbf{VCU} \equiv \mathbf{HVCU}$.

We describe below the strategy of our proof, so that the basic line of thought is transparent from the beginning.

Proof strategy Not all *VC* models can be Hammingized by simply manipulating their valuations (no need to say preserve the truth of some φ), but any *VC* model which has some tree structure can. Moreover, (Friedman and Halpern 1994) offered a tree construction from some *VC* model while preserving the truth of some formula φ. Hence, we aim to Hammingize the Friedman-Halpern tree *VC* model while not affecting the truth of φ. To divide the proof into steps and conquer them separately, we will show that if φ is satisfiable in **VC** then it is satisfied in a pointed HVC model (M, w_0) which fulfills the following missions:

1. Hammingianization: M induces an HVC model M';
2. Truth-preservation: $(M', w_0) \models \varphi$.

The strategy for *VCU* is the same but need one further treatment.

4.1 A Failed Attempt

Let us start with an easy but failed attempt.

A simple thought for Mission 1, Hammingianization, is to keep the worlds and their similarity relations, and *only* manipulate the valuation on $Atm \setminus atm(\varphi)$, resulting in a new

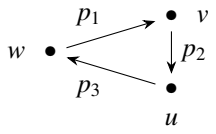

Figure 2. The vicious circle in Example 1. Arrows denote the relevant selection functions. E.g., $\sigma_w(p_1) = \{v\}$ due to the fact that $v \prec_w u$, though p_3 is in both $V(v)$ and $V(u)$.

valuation V' so that the Hamming distance by V' is in accordance with the similarity relations in the original model. We may call such a VC model "substantially Hammingian".[8]

Definition 12 (Substantial Hammingianness)**.** *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a VC model. We call M* substantially Hammingian *if there is a valuation V', s.t. $M' = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V')$ is a Hammingian model.*

Naturally we ask the following question: Are all VC models substantially Hammingian? The answer is, however, negative, whenever a VC model has a vicuous circle.

Definition 13 (Vicious circle)**.** *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a VC model. We say that M has a vicious counterfactual circle, if $\exists w_0, w_1, \ldots, w_n \in W$ s.t. $w_0 \preceq_{w_1} w_2 \ldots$, and $w_{n-2} \preceq_{w_{n-1}} w_n$, and $w_{n-1} \preceq_{w_n} w_0$, but $w_n \prec_{w_0} w_1$.*

Then we have the following impossibility result.

Proposition 4. *Any VC model that has a vicious circle is not substantially Hammingian.*

Proof. We show the case when the circle consists of 3 worlds; the other cases are similar. Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a VC model and $w, v, u \in W$ form a vicious circle. For whatever V', we should have $\hbar_{V'}(w, v) \leq \hbar_{V'}(v, u) \leq \hbar_{V'}(u, w) < \hbar_{V'}(w, v)$, by the definition of Hammingian model and using the condition of the circle. But $\hbar_{V'}(w, v) = \hbar_{V'}(v, w)$, a contradiction. □

The same definitions and same result, as its proof indicates, apply to VCU models as well.

Example 1. *Let $\varphi^\dagger$ be the formula $\neg p_1 \wedge (p_1 \boxdot\!\!\rightarrow (\neg p_2 \wedge (p_2 \boxdot\!\!\rightarrow (\neg p_3 \wedge (p_3 \boxdot\!\!\rightarrow p_1))))$. Let a VCU model $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be s.t. $W = \{w, v, u\}$, $V(w) = \{p_2, p_3\}, V(v) = \{p_1, p_3\}, V(u) = \{p_1, p_2\}$, and $v \prec_w u, u \prec_v w, w \prec_u v$. This is depicted in Figure 2. Then $(M, w) \models \varphi^\dagger$, but M is not substantially Hammingian.*

Actually this is a difficulty not only to Hamming distance, but any total order on pairs of worlds which intends to extend the sets of triple relations on worlds. We can take advantage of the study in (Williamson 1988), which helps us prove the following proposition.

Proposition 5. *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a VCU model. It has no vicious circle, if and only if there exists a total order $\leq$ on W^2 s.t. if $v \preceq_w u$ then $(v, w) \leq (w, u)$.*

[8]Let us distinguish "substantially" and "potentially" Hammingian. The former only needs to manipulate V to become Hammingian; while the later may copy worlds to unravel the vicious circle, as we will do. Actually all VC models are potentially Hammingianizable by first transforming to substantially Hammingian ones.

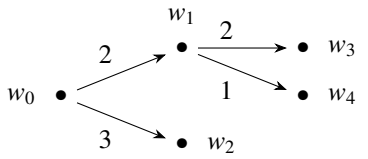

Figure 3. Given such a weighted tree, one can build an HV model as follows: let W consist of the five vertices, and take a V s.t. $\forall w_i, w_j, \hbar_V(w_i, w_j) = n$ if $\pi(w_i, w_j)$ is weighted n. Restricting $W_{w_i} = W$ for all w_i the model is HVCU, otherwise HVC. But of course, we cannot guarantee any truth-preservation at this stage.

Proof sketch. Necessity is shown by Example 1. For sufficiency, we need constitute a total ordering $\leq$ on W^2. First, notice that every $\preceq_w$ can be seen as a partial ordering on W^2 by stipulating $\forall v \in W_w, u \notin W_w, v \prec_w u$. Hence we union all of them to obtain a partial ordering on W^2, noted $\preceq$. No vicious circle ensures the union. The second step is taking transitive closure of it, noted $\preceq'$. The next step is taking the quotient w.r.t. equivalence relation of $\preceq'$, i.e. the relation on the equivalent classes of W^2 with respect to $\preceq'$, noted $\preceq''$. Last, we can use the famous theorem in order theory that every partial order can be extended to a total order. □

4.2 Weighted Tree is Hammingian

What can we learn from the former failure? The lesson is that without further constraint on the original VC or VCU model, the ternary relations may conflict with each other.

The last proposition indicates that the necessary condition of being a Hammingian model is no vicious circle. So the tree structure appears as a natural choice.

The intuition illustrated in Figure 3 is that if the model associates with a tree structure, and moreoever the tree is weighted, then it can be Hammingized by adding the weights of edges of the path between any two vertices (worlds). But before formalizing the intuition, let us make two remarks.

1) We beg patience at this stage about where the mysterious tree structure of a model comes from. It will be clear in the next steps.

2) We recall the basic notions of graph theory. For any two points (vertices) w, v, (w, v) denotes the undirected edge between w and v. A path between w and v is a sequence of vertices $(w_1, \ldots, w_n)$ s.t. $w = w_1, v = w_n$ and (w_i, w_{i+1}) is an edge for $1 \leq i < n$. A tree is an undirected graph where each two vertices w and u have exactly one path, denoted by $\pi(w, v)$. We write $(w_i, w_j) \in \pi(w, v)$ if (w_i, w_j) is a member of the sequence.

A weighted tree is a triple $G^\# = (W, E, \#)$ where $G = (W, E)$ is a tree (with $E \subseteq W \times W$) and $\# : E \longrightarrow \mathbb{N}$. The weight of a path $\pi(w, v)$ in $G^\#$ is $\#\pi(w, v) =_{def} \sum_{(w_i,w_j)\in\pi(w,v)} \#(w_i, w_j)$.

Definition 14 (Weighted tree VC model)**.** *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a VC model, for which there exists an associated weighted tree $G^\# = (W, E, \#)$, s.t. $\forall w \in W$, $\forall v, u \in W_w$, $v \preceq_w u \iff \#\pi(v, w) \leq \#\pi(w, u)$.*

Lemma 1. *Let $M = (W, (\preceq_w)_{w\in W}, V)$ be a finite VC model associated with a weighted tree $G^{\#}$. Then, there is an HVC model $M' = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V')$, s.t. $\forall w, v \in W$, $\hbar_{V'}(w, v) = 2 \times \#\pi(w, v)$, and $\forall w \in W$, $|V(w)|$ is finite.*

Proof. Since no formula need be truth-preserved here, we construct V' ignoring V. We take a series of disjoint unions $X_1 \cup Y_1 \cup X_2 \cup Y_2 \cup \ldots X_{|E|} \cup Y_{|E|} \subset Atm$ and enumerate E as $e_1, e_2, \ldots, e_{|E|}$, s.t. $|X_i| = |Y_i| = \#(e_i)$ for all $1 \leq i \leq |E|$. For every p that is not in those disjoint unions, let $p \notin V'(w)$ for all w. The construction of V' says for all $e_i = (w_j, w_k) \in E$, if $\pi(w_0, w_j) \subset \pi(w_0, w_k)$, viz. w_j is nearer to w_0 than w_k, then let $V'(w_j) \cap (X_i \cup Y_i) = X_i$ and $V'(w_k) \cap (X_i \cup Y_i) = Y_i$; for all e_l not linking w_j, simply let $V(w_j) \cap (X_l \cup Y_l) = \emptyset$. Thus, for any $V'(w), V'(v)$ they differ on $2 \times \#\pi(w, v)$ many variables, which makes the desired $\hbar_{V'}(w, v) = 2 \times \#\pi(w, v)$. $|V(w)|$ is finite because $G^{\#}$ has finitely many edges with finite weights. □

The proposition below directly follows from the lemma.

Proposition 6. *All weighted tree VC models are substantially Hammingian.*

4.3 VC ≡ HVC

Before exhausting the reader's patience, we now reveal where the tree comes from: *it is constructed according to subformulas in the formula φ, noted sub(φ), of interest*. The construction is described in (Friedman and Halpern 1994).

Proposition 7 ((Friedman and Halpern 1994))**.** *If φ is satisfiable in* **VC***, then φ is satisfiable in some tree VC model.*

The proof relies on a series of lemmas to construct such a tree, which we shall call the *FH tree* after the authors. For the sake of both self-containedness and simplicity, we rephrase how the tree is constructed.

Friedman-Halpern tree for VC model The first key notion is $basic_i(\varphi) \subseteq atm(\varphi) \cup sub_{\Box\!\!\rightarrow}(\varphi)$ where $sub_{\Box\!\!\rightarrow}(\varphi)$ denotes the subformulas of φ whose principal connective is $\Box\!\!\rightarrow$. Intuitively, $basic_i(\varphi)$ is defined as the union of all atoms in φ and counterfactuals in *exactly* the i-th level of the nesting of φ. A formal definition is:

$$basic_i(p) = \begin{cases} \{p\} & \text{if } i = 0, \\ \emptyset & \text{otherwise;} \end{cases}$$

$$basic_i(\neg\varphi) = basic_i(\varphi);$$

$$basic_i(\varphi \wedge \psi) = basic_i(\varphi) \cup basic_i(\psi);$$

$$basic_i(\varphi \Box\!\!\rightarrow \psi) = \begin{cases} \{\varphi \Box\!\!\rightarrow \psi\} & \text{if } i = 0, \\ basic_{i-1}(\varphi) \cup basic_{i-1}(\psi) & \text{otherwise.} \end{cases}$$

Take $\varphi^\dagger = \neg p_1 \wedge (p_1 \Box\!\!\rightarrow (\neg p_2 \wedge (p_2 \Box\!\!\rightarrow (\neg p_3 \wedge (p_3 \Box\!\!\rightarrow p_1))))$ from Example 1, then $basic_0(\varphi^\dagger) = \{p_1, p_1 \Box\!\!\rightarrow (\neg p_2 \wedge (p_2 \Box\!\!\rightarrow (\neg p_3 \wedge (p_3 \Box\!\!\rightarrow p_1))))\}$, $basic_1(\varphi^\dagger) = \{p_1, p_2, p_2 \Box\!\!\rightarrow (\neg p_3 \wedge (p_3 \Box\!\!\rightarrow p_1))\}$, $basic_2(\varphi^\dagger) = \{p_2, p_3, p_3 \Box\!\!\rightarrow p_1\}$ and $basic_3(\varphi^\dagger) = \{p_1, p_3\}$.

We describe an FH tree given a finite VC model $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ and a formula φ s.t. $(M, w_0) \models \varphi$. The tree iteratively "chooses" worlds in W as vertices according to vertices and formulas at the previous level. The root is w_0. Since the function of choosing is not necessarily injective, for any vertex v we write v^{-1} for the chosen world in M. But we only write w_0 for simplicity. Level 0 has only the root w_0. At level 1, for any $\xi \Box\!\!\rightarrow \theta \in basic_0(\varphi)$ there is a vertex named as $w_{0,\xi\Box\!\!\rightarrow\theta}$. And $w^{-1}_{0,\xi\Box\!\!\rightarrow\theta}$ was chosen from M with the following constraints:

1. $w^{-1}_{0,\xi\Box\!\!\rightarrow\theta} \in \sigma_{w_0}(\xi \Box\!\!\rightarrow \theta)$, if $(M, w_0) \models \xi \Box\!\!\rightarrow \theta$ and $\sigma_{w_0}(\xi \Box\!\!\rightarrow \theta) \neq \emptyset$;
2. $w^{-1}_{0,\xi\Box\!\!\rightarrow\theta}$ is w_0, if $\sigma_{w_0}(\xi \Box\!\!\rightarrow \theta) = \emptyset$;
3. $w^{-1}_{0,\xi\Box\!\!\rightarrow\theta} \in \sigma_{w_0}(\xi \Box\!\!\rightarrow \theta)$ and $(M, v) \models \neg\theta$, if $(M, w_0) \models \neg(\xi \Box\!\!\rightarrow \theta)$.

Notice that when $(M, w_0) \models \xi$, $w^{-1}_{0,\xi\Box\!\!\rightarrow\theta}$ has to be w_0. Naturally, for every such vertex, we draw an edge between it and w_0 to obtain a (sub)tree. Then define a model $M^{w_0} = (W^{w_0}, (W^{w_0}_v)_{v\in W_0}, (\preceq^{w_0}_w)_{w\in W_{w_0}}, V^{w_0})$ s.t. $v \in W^{w_0}$ if (w_0, v) is an edge; $V^{w_0}(v) = V(v^{-1})$. Then, we simply put $W^{w_0}_v = \emptyset$ if $v \neq w_0$. Let $\preceq^{w_0}_{w_0}$ be a total order on W^{w_0} s.t. **Centering** is satisfied and $\forall w_{0,\xi\Box\!\!\rightarrow\theta}, w_{0,\xi'\Box\!\!\rightarrow\theta'} \in W^{w_0}$, $w_{0,\xi\Box\!\!\rightarrow\theta} \preceq^{w_0}_{w_0} w_{0,\xi'\Box\!\!\rightarrow\theta'}$ iff $(M, w_0) \models \xi \vee \xi' \Box\!\!\rightarrow \xi$.

Now let v be a vertex at level 1. We define $\varphi_v := \bigwedge_{\psi\in basic_1(\varphi),(M,v^{-1})\models\psi} \psi \wedge \bigwedge_{\psi\in basic_1(\varphi),(M,v^{-1})\models\neg\psi} \neg\psi$. We recursively apply the procedure on (M, v^{-1}) and φ_v to obtain a subtree and a submodel. Since $sub_{\Box\!\!\rightarrow}(\varphi_v) \subseteq sub_{\Box\!\!\rightarrow}(\varphi), basic_i(\varphi_v) \subseteq basic_{i+1}(\varphi)$, the construction terminates. Finally, we union all of them to obtain the FH tree and its associated VC model noted M^t. Figure 4 illustrates a tree truth-preserving φ relative to VC.[9]

Remark & convention For readability we save the recursively defined function for the "standard name" of worlds in M^t, which takes the form $w_{*,\xi\Box\!\!\rightarrow\theta}$ where $\xi \Box\!\!\rightarrow \theta \in basic_i(\varphi)$ for some i and w_* is the name of a world at level i. We may enumerate these names and hence $w_k = w_{j,\psi}$ if w_k exists w.r.t. some world w_j and formula $\xi \Box\!\!\rightarrow \theta$. Notice that while $w_j \neq w_k$ for $j \neq k$, it is possible that $w_j^{-1} = w_k^{-1}$.

Example 2. *Figure 4 is the graph G of the tree model M^t constructed from some finite VC model M and $(p \Box\!\!\rightarrow (q \Box\!\!\rightarrow r) \wedge \neg((q \Box\!\!\rightarrow q) \Box\!\!\rightarrow r)$. The formula attached to every edge denotes the member of $basic_i(\varphi)$ making the target vertex exist. For example, the arrow with p means that w_1 is chosen from $\sigma_{w_0}(p \Box\!\!\rightarrow (q \Box\!\!\rightarrow r))$ during the tree construction. Formulas in each world v are the conjuncts of φ_v as defined in the proof of Proposition 7. Notice, e.g. though w_4 exists for sake of $\neg(q \Box\!\!\rightarrow p)$, we also know $(M^t, w_4) \models r$, because w_4 is chosen from M as $w_4^{-1} \in \sigma_{w_1^{-1}}(q)$. Since $(M, w_1^{-1}) \models q \Box\!\!\rightarrow r$, it must be $r \in V^t(w_3)$.*

Lemma 2. *Let $(M, w_0) \models \varphi$ where M is a finite VC model. Then there is an FH tree VC model M^t built from (M, w_0) and φ, s.t. $(M^t, w_0) \models \varphi$.*

[9] We defined our tree as undirected in accordance with the semantics of $\preceq_w$. In fact, the tree in the construction above is better understood as directed in accordance with the semantics of σ_w. However, since it causes minor problems of understanding, we do not add more definitions to increase the opaqueness.

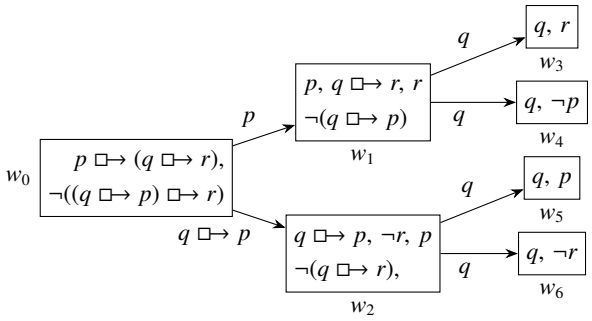

Figure 4. Example for $(p \Box\!\!\rightarrow (q \Box\!\!\rightarrow r) \wedge \neg((q \Box\!\!\rightarrow q) \Box\!\!\rightarrow r)$

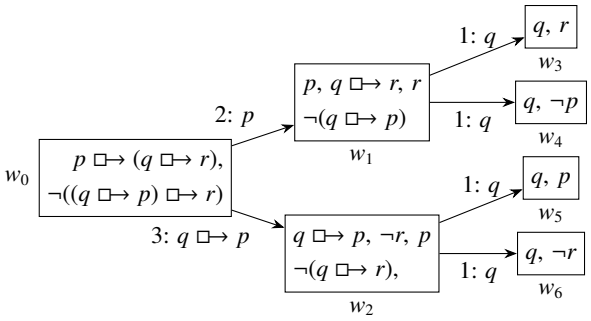

Figure 5. A weighted tree $G^{\#}$ for G in Figure 4

Hammingize the tree VC model Are we done? Almost yet not. We know that FH tree construction is truth-preserving, and a weighted tree is substantially Hammingian. It remains to Hammingize the weighted FH tree in order to turn Thesis 1 into a theorem.

Theorem 1. *Let Atm be infinite. Then* **VC** ≡ **HVC**.

Proof. **HVC** ⊂ **VC**, so only need prove the only-if-part. For any φ satisfiable in **VC**, using the filtration result of (Segerberg 1989), there is a finite model $M^f = (W^f, (W^f_w)_{w\in W^f}, (\preceq^f_w)_{w\in W^f}, V^f)$ s.t. $(M^f, w_0) \models \varphi$, and particularly $\bigcup_{v\in W^f} V^f(v) \subseteq atm(\varphi)$. That is, no world in W^f verifies any variable outside of $atm(\varphi)$. We build an FH tree model $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ from (M^f, w_0) and φ. By Lemma 2, $(M, w_0) \models \varphi$. A weighted tree $G^{\#} = (W, E, \#)$ is defined s.t. $E = \{(w_j, w_{j,\xi\Box\!\!\rightarrow\theta}) : w_j, w_{j,\xi\Box\!\!\rightarrow\theta} \in W)\}$, and $\forall w, v, u \in W$, if $v \preceq_w u$ then $\#(v, w) \leq \#(w, u)$. We obtain V' by assembling three valuations. The first one is some V^h which enables Hammingization in Lemma 1, with $\bigcup_{v\in W} V^h(v) \cap atm(\varphi) = \emptyset$. The second one is V, because we want $V'(w) \cap atm(\varphi) = V(w) \cap atm(\varphi)$ for truth-preserving φ. But now the Hamming distance perturbs, which needs a third one V^b to "counterbalance" it. Let $w^\circ, w_\circ$ be two worlds that differ at most on atoms in $atm(\varphi)$, say, $\hbar_V(w^\circ, w_\circ) = n$. Let $\bigcup_{v\in W} V^b(v) = X$ be disjoint from $atm(\varphi)$ and $\bigcup_{v\in W} V^h(v)$ with $|X| = n$. Then we enumerate $V(w^\circ) \triangle V(w_\circ)$ as $p_1, p_2, \ldots p_n$, and X as $q_1, q_2, \ldots q_n$,

1. let $V^b(w^\circ) \cap X = X$ and $V^b(w_\circ) \cap X = \emptyset$;
2. $\forall v \in W$, $\forall p_i \in atm(\varphi)$, $p_i \in V(v) \cap V(w^\circ)$ if and only if $q_i \notin V^b(v)$.

This step guarantees that $\forall w, v \in W, |V^b(w) \cap (X \cup atm(\varphi))| = |V^b(v) \cap (X \cup atm(\varphi))|$. Namely w and v are "numerically equal" regarding $atm(\varphi) \cup X$, so that V^h can do its job right. Finally we let $\forall w \in W, V'(w) = V(w) \cup V^b(w) \cup V^h(w)$. Obviously, M' is Hammingian and still $(M', w_0) \models \varphi$. □

4.4 VCU ≡ HVCU

We did not apply the method above directly to VCU, because of an apparent shortcoming and a potential danger. 1) The tree VC model is not uniform but "local": each W_v contains only v and its adjacents. We can of course extend $\preceq_v$ to obtain **Uniformity** by the information of the weighted FH tree. But then 2) one may suppose that $(M_v, v) \models \neg\xi \Box\!\!\rightarrow \bot$ holds vacuously, i.e. $\sigma^v_v(\neg\xi) = \emptyset$, but after the extending possibly in M, $\sigma_v(\neg\xi) \neq \emptyset$. Hence finally $(M, v) \models \neg(\neg\xi \Box\!\!\rightarrow \bot)$.[10]

We are going to show that refining the tree construction in a certain way, not only the shortage is overcome, the potential danger is actually no danger. The key fact is that, though every M_{w_j} is constructed "shortsightedly", the original finite model this time satisfies **Uniformity**. Hence if some $\xi \Box\!\!\rightarrow \bot$ is vacuously true in the tree model, that must be already vacuously true in the original model.

Instead of first presenting the refined FH tree for a VCU model and then Hammingianize it, for simplicity we do the two steps simultaneously.

Forward-weighted tree for HVCU model Let the FH tree construction remain the same, but the input model is VCU instead of VC. The output model is (still) a tree VC model. Now our goal is an HVCU model, where **Uniformity** is obtained naturally by generalizing all W^t_w to the whole W^t; and the information of the weighted tree truth-preserves certain formulas.

To this end we need ensure the distance between worlds to go with the "direction" of the tree construction, so that for each v, its sets of closest worlds regarding $basic_0(\varphi_v)$ remain invariant. Thus we need the wanted weighted tree to have a particular global property defined as following.

Definition 15 (Forward-weighted tree)**.** *Let M be a tree VC model associated with a weighted tree $G^{\#} = (W, E, \#)$ and w_0 be the root. We call $G^{\#}$* forward-weighted*, if $\forall w, v, u$, if $\pi(w_0, w) \subset \pi(w_0, v) \subset \pi(w_0, u)$, then $\#\pi(w, v) > \#(v, u)$.*

In plain words, the farther we go from the root, the smaller weights we assign the edges. If in the graph (w_0, w, v, u) forms a path, then $\#(w, v) > \#(v, u)$. So when we search the closest worlds of v regarding some ξ, we will not "go back" to w. This is the intuition where the term comes from.

It is not hard to see, similar to what we did in the last subsection, that a forward-weighted tree associating a VCU model induces an HVCU model. In particular, we have the following lemma, which is proven similar to Theorem 1.

[10](Friedman and Halpern 1994) mentioned a similar concern in satisfiability problem in **VCU** and hinted a solution which raises the complexity to EXPTIME. They claimed to leave the details to the full paper. According to personal communication, no full paper.

Example 3. *Figure 5 hammingizes the G in Figure 4 as $G^{\#}$. For example, the edge labelled 2:p is because we want $w_1 \in \sigma_{w_0}(p)$. Notice if the edge with $q \boxrightarrow p$ weighted as 2, then we would also have $w_2 \in \sigma_{w_0}(p)$ since $(M, w_2) \models p$, which would make $(M, w_0) \not\models p \boxrightarrow (q \boxrightarrow r)$, since $(M, w_2) \models \neg(q \boxrightarrow r)$. Also, if $(M, w_0) \models q \wedge \neg r$ and we weighted the edge of $r \boxrightarrow q$ as 1, then we would have $(M, w_2) \not\models q \boxrightarrow r$, since $w_0 \in \sigma_{w_2}(q)$, and we "go back". Finally, construct an HV model M' via $G^{\#}$ instructed by Theorem 1, s.t. for its V':*

- $\hbar_{V'}(i, j) = 2n$ *iff* $\#\pi(w_i, w_j) = n$;
- $\forall p_k \in \{p, q, r\}, p_k \in V'(w_i)$ *iff* $p_k \in V(w_i)$.

Lemma 3. *Let $M^t = (W^t, (W^t_w)_{w\in W^t}, (\preceq^t_w)_{w\in W}, V)$ be an FH tree VC model constructed from some finite VCU model and φ. Then there exists a forward-weighted tree $G^{\#}$ of M, which induces an HVCU model $M' = (W^t, (W'_w)_{w\in W}, (\preceq'_w)_{w\in W}, V')$, s.t. $\forall w \in W$, $W'_w = W^t$ and $V(w) \cap atm(\varphi) = V'(w) \cap atm(\varphi)$.*

Key lemma We have enabled FH tree VC model with **Uniformity** and Hammingized it to obtain an HVCU model. We aim to show a key lemma to disprove the potential danger.

Lemma 4. *Let $M = (W, (W_w)_{w\in W}, (\preceq_w)_{w\in W}, V)$ be a finite VCU model, $w_0 \in W$, $\varphi \in \mathcal{L}(Atm)$ s.t. $(M, w_0) \models \varphi$. Let $M' = (W^t, (W^t_w)_{w\in W^t}, (\preceq'_w)_{w\in W}, V')$ be an HVCU model constructed through FH tree and Lemma 1. Then $(M, w_0) \models \varphi$ if and only if $(M', w_0) \models \varphi$.*

Proof. Inherited from the FH tree for VC models, we have $\forall w_j \in W^t$ chosen by the tree construction at level i, $\forall \psi \in basic_i(\varphi)$, $(M, w_j^{-1}) \models \psi$ if and only if $(M', w_j) \models \psi$.

So the only concern is when $\psi \in sub(\varphi)$ is some $\chi \boxrightarrow \chi'$ such that either χ is vacuously true at (M, w_j^{-1}) but not vacuously true at (M', w_j); or the other way around. Notice ψ may not be at the same level as w_j is chosen, but crucially it must be $\psi \in basic_k(\varphi)$ for some k. For convenience instead of saying vacuously true or $\sigma_{w_j^{-1}}(\chi) = \emptyset$ we write $(M, w_j^{-1}) \models \chi \boxrightarrow \bot$. We do induction on the conditional degree of χ.

The induction basis is $cd(\chi) = 0$, viz. χ is Boolean. If $(M, w_j^{-1}) \models \chi \boxrightarrow \bot$, then $\forall v \in W^t$ we have $(M, v^{-1}) \models \neg\chi$. Since χ is Boolean we have $(M', v) \models \neg\chi$, hence $(M', w_j) \models \chi \boxrightarrow \bot$. For the other direction let $(M', w_j) \models \chi \boxrightarrow \bot$, and we need show $(M, w_j^{-1}) \models \chi \boxrightarrow \bot$. Notice, crucially, that $\chi \boxrightarrow \chi'$ must occur in $basic_k(\varphi)$ for some k. At level k we must choose a world w_l to decide whether $\chi \boxrightarrow \chi'$ holds at (M', w_l) according to what happens at (M, w_l^{-1}). But whatever $w_l^{-1} \in W$ is, it must be $(M, w_l^{-1}) \models \chi \boxrightarrow \bot$, otherwise at level $k+1$ we would have chosen a $w_{l,\chi\boxrightarrow\chi'}$ s.t. $(M', w_{l,\chi\boxrightarrow\chi'}) \models \chi$, which eventually made $(M', w_j) \models \neg(\chi \boxrightarrow \bot)$, a contradiction. This indicates $\forall v \in W, (M, v) \models \neg\chi$. Since χ is Boolean, $\forall w \in W^t, (M, w) \models \neg\chi$, viz. $(M, w_j^{-1}) \models \chi \boxrightarrow \bot$ as we want.

Now we run the induction. Suppose for any subformula of conditional degree n, it is true at (M, v^{-1}) if and only if true at (M', v) for all $v \in W^t$. Now we consider χ with $cd(\chi) = n+1$ and show $(M, w_j^{-1}) \models \chi \boxrightarrow \bot$ iff $(M', w_j) \models \chi \boxrightarrow \bot$. It needs a further induction on the main connective of χ.

1) The case of conjunction is straightforward. 2) If χ has the form $\xi \boxrightarrow \theta$ and $(M, w_j^{-1}) \models (\xi \boxrightarrow \theta) \boxrightarrow \bot$, then suppose towards a contradiction that $\exists v \in W^t$ s.t. $(M', v) \models \xi \boxrightarrow \theta$. By induction hypothesis we have $(M, v^{-1}) \models \xi \boxrightarrow \theta$, which contradicts $(M, w_j^{-1}) \models (\xi \boxrightarrow \theta) \boxrightarrow \bot$. For the other direction suppose towards a contradiction that $(M', w_j) \models (\xi \boxrightarrow \theta) \boxrightarrow \bot$ but $(M, w_j^{-1}) \models \neg((\xi \boxrightarrow \theta) \boxrightarrow \bot)$. Notice, crucially, that $(\xi \boxrightarrow \theta) \boxrightarrow \chi'$ occurs in $basic_k(\varphi)$ for some k. Thus at level k of the tree construction there was a $w_l \in W^t$ which chose a $w_{l,(\xi\boxrightarrow\theta)\boxrightarrow\chi'}$ from W for the level $k+1$ according to whether $(\xi \boxrightarrow \theta) \boxrightarrow \chi'$ holds at (M, w_l^{-1}). It must be $(M, w_l^{-1}) \models \neg((\xi \boxrightarrow \theta) \boxrightarrow \bot)$ because of the supposition $(M, w_j^{-1}) \models \neg((\xi \boxrightarrow \theta) \boxrightarrow \bot)$. Thus the tree construction chose a $w_{l,(\xi\boxrightarrow\theta)\boxrightarrow\chi'}$ s.t. $(M, w^{-1}_{l,(\xi\boxrightarrow\theta)\boxrightarrow\chi'}) \models \xi \boxrightarrow \theta$. By induction hypothesis $(M', w_{l,(\xi\boxrightarrow\theta)\boxrightarrow\chi'}) \models \xi \boxrightarrow \theta$, contradicting $(M', w_j) \models (\xi \boxrightarrow \theta) \boxrightarrow \bot$ as we want.

3) If χ has the form $\neg\zeta$, we need a further induction. But the only interesting case is when χ equals some $\neg(\xi \boxrightarrow \theta)$. Assume $(M, w_j^{-1}) \models \neg(\xi \boxrightarrow \theta) \boxrightarrow \bot$, we need now show $(M', w_j) \models \neg(\xi \boxrightarrow \theta) \boxrightarrow \bot$. Suppose not towards a contradiction. Then $\exists v \in W^t, (M', v) \models \neg(\xi \boxrightarrow \theta)$, viz. $\exists u \in \sigma^t_v(\xi), (M', u) \models \xi \wedge \neg\theta$. By induction hypothesis, $(M, u^{-1}) \models \xi \wedge \neg\theta$. By **Centering** we have $(M, u^{-1}) \models \neg(\xi \boxrightarrow \theta)$, contradicting the assumption.

For the other direction, suppose towards a contradiction that $(M', w_j) \models \neg(\xi \boxrightarrow \theta) \boxrightarrow \bot$ but $(M, w_j^{-1}) \models \neg(\neg(\xi \boxrightarrow \theta) \boxrightarrow \bot)$. Now notice, crucially, that $\neg(\xi \boxrightarrow \theta) \boxrightarrow \chi'$ occurs in $basic_k(\varphi)$ for some k. Then at level k there was a $w_l \in W^t$ which chose a $w_{l,\neg(\xi\boxrightarrow\theta)\boxrightarrow\chi'}$ from W for the level $k+1$. Because of the eventual $(M', w_j^{-1}) \models \neg(\xi \boxrightarrow \theta) \boxrightarrow \bot$ it must be during the tree construction we had $(M', w_{l,\neg(\xi\boxrightarrow\theta)\boxrightarrow\chi'}) \models \neg(\xi \boxrightarrow \theta)$. By induction hypothesis $(M, w^{-1}_{l,\neg(\xi\boxrightarrow\theta)\boxrightarrow\chi'}) \models \neg(\xi \boxrightarrow \theta)$, a wanted contradiction. □

Now Thesis 2 becomes a theorem.

Theorem 2. *Let Atm be infinite. Then* **VCU** ≡ **HVCU**.

Proof. Since **HVCU** ⊂ **VCU**, we only need prove the rest direction. Similar to the proof of Theorem 1, for any φ satisfiable in **VCU**, we start with a filtration model $M^f = (W^f, (W^f_w)_{w\in W^f}(\preceq^f_w)_{w\in W^f}, V^f)$ and a $w_0 \in W^f$ s.t. $(M^f, w_0) \models \varphi$. Then we build an FH tree model M by Lemma 3. The next step is associating M with a forward-weighted tree $G^{\#}$ and construct an HVCU model M'. Finally, with the help of Lemma 4, an induction on φ can show that $(M', w_0) \models \varphi$, which is what we want. □

5 Application to Classifier Explanation

In this section, we are going to show how the Hammingian semantics for the logic of conditionals can be used to define an interesting notion of counterfactual explanation. This notion has been widely discussed in the area of explainable AI (XAI) (Mittelstadt, Russell, and Wachter 2019; Mothilal, Sharma, and Tan 2020; Sokol and Flach 2019; Kenny and Keane 2021). It is paramount to explaining the decisions of classifier systems. Here, we focus on *binary* classifier systems. We define counterfactual explanation by

the following abbreviation:

$$\mathsf{CfXp}(\varphi, \psi) =_{def} \psi \wedge (\neg\varphi \boxdot\!\!\rightarrow \neg\psi).$$

$\mathsf{CfXp}(\varphi, \psi)$ has to be read "the fact φ counterfactually explains the fact ψ". The latter just means that ψ is true and that if φ was false, ψ would be false as well. Formula φ is the *explanans*, while ψ is the *explanandum*.

We illustrate the definition with the help of a concrete example of a binary classifier that has to decide whether an application for a loan to the bank has to be accepted (acc) or rejected ($\neg acc$). The classifier's decision depends on the values of three binary variables: whether the applicant has a permanent job (pe), whether she/he earns a salary of at least 3000€ per month (sa), and whether she/he is a European citizen (eu). For the sake of modeling, we suppose that the set of atomic propositions Atm_0 is finite.

The binary classifier is fully described in Table 2. It is represented by a formula of our language, namely, the $\boxdot$-modality followed by the canonical DNF of the boolean function corresponding to it:

$$\begin{aligned}\varphi_{cl} =_{def} \boxdot(acc \leftrightarrow ((pe \wedge \neg sa \wedge \neg eu) \vee \\ (\neg pe \wedge sa \wedge eu) \vee \\ (pe \wedge \neg sa \wedge eu) \vee \\ (pe \wedge sa \wedge \neg eu) \vee \\ (pe \wedge sa \wedge eu)).\end{aligned}$$

The following formula characterizes completeness of the model, that is, the fact that all possible valuations of propositional atoms are in it:

$$\varphi_{comp} =_{def} \bigwedge_{X \subseteq Atm_0} \Diamond(\bigwedge_{p \in X} p \wedge \bigwedge_{q \in Atm_0 \setminus X} \neg q).$$

The assumption that the set Atm_0 is finite is essential. Otherwise φ_{comp} woud be an infinite formula. We have the following validity:

$$\models_{\mathbf{HVCU}} (\varphi_{comp} \wedge \varphi_{cl} \wedge (\neg pe \wedge \neg sa \wedge eu)) \rightarrow \mathsf{CfXp}(\neg sa, \neg acc).$$

This means that, under the completeness assumption φ_{comp} and given the classifier described by the formula φ_{cl} if the applicant has no permanent job, her/his salary is lower than 3000€ per month and she/he is European, then the fact of not having a monthly salary of at least 3000€ counterfactually explains the failure of her/his application. Indeed, under the hypothesis that the applicant has the features $\neg pe$, $\neg sa$ and eu, if her/his salary was at least 3000€ per month, her/his application would be successful.

The requirement that Atm is finite is at odds with our hypothesis that there is an infinite reserve of propositional variables, which was instrumental in proofs of Theorems 1 and 2. Instead of the **HVCU** validity checking one could use symbolic model checking and suppose that the set of possible valuations of an HVCU model is characterized by a propositional formula χ. Then models (M, w) can be replaced by pairs (χ, v) where χ is a propositional formula and $v \subseteq Atm$ is a valuation satisfying χ in propositional logic. It remains to define an algorithm checking whether a formula φ is satisfied by a pair (χ, v). In the case of the present example φ is $\mathsf{CfXp}(\neg sa, \neg acc)$, χ is φ_{cl} without the box-operator and v is $\{eu\}$. We conjecture that symbolic model checking so formulated is PSPACE-complete. The upper bound is provable by giving a PSPACE algorithm, while the lower bound via a reduction of TQBF into our problem.

Permanent job	>3000€ salary	EU citizen	Accept
0	0	0	No
0	0	1	No
0	1	0	No
1	0	0	Yes
0	1	1	Yes
1	0	1	Yes
1	1	0	Yes
1	1	1	Yes

Table 2. A classifier for loan application

6 Conclusion and Further Discussion

We studied the Hammingian V models, and in particular proved **VC** ≡ **HVC** and **VCU** ≡ **HVCU** given infinite variables in the language. Notice the precondition, which is because Hammingization relies on manipulating variables out of $atm(\varphi)$. This cannot happen without an *unbounded* number of fresh variables: it is known that when Atm is *finite* then Hamming distance can be axiomatized (Liu and Lorini 2023) by, essentially, taking the conjunction of a maximal consistent set of literal to express a state syntactically.

The technical conclusion is that the property of being Hammingian is unaxiomatizable given the basic language of counterfactuals with infinite variables. The most straightforward philosophical interpretation is that any abstract notion of distance, e.g. epistemic entrenchment, system of spheres, etc, can be re-interpreted/implemented by Hamming distance by means of "hidden variables". In this sense we call Hamming distance "grounded" for VC and VCU.

For future work, Friedman & Halpern (Friedman and Halpern 1994) claimed that the satisfiability problem for **VCU** is EXPTIME-complete. They gave a PSPACE algorithm *check-tree* for V models without **Uniformity**, and mentioned that it does not work directly for models with **Uniformity** but needs possibly exponential expansion. But since our study on satisfiability showed that FH trees apply to VC and VCU models in a similar way, it is interesting to check that whether the complexity can be lower. Also we will address the conjecture of the PSPACE-complete complexity of model checking at the end of Section 5.

Another intriguing topic is that given that Hamming distance grounds the comparative similarity when it is a total preorder, does another concrete definition of distance, subset relation of valuations, ground the partial order version?

Acknowledgements

This work is supported by the EU ICT-482020 project TAILOR (No.952215) and by the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI).

References

Aiguier, M.; Atif, J.; Bloch, I.; and Hudelot, C. 2018. Belief revision, minimal change and relaxation: A general framework based on satisfaction systems, and applications to description logics. *Artificial Intelligence* 256:160–180.

Burgess, J. P. 1981. Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic* 22(1):76–84.

Dalal, M. 1988. Investigations into a theory of knowledge base revision: preliminary report. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, volume 2, 475–479.

Darwiche, A., and Hirth, A. 2020. On the reasons behind decisions. In *24th European Conference on Artificial Intelligence (ECAI 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 712–720. IOS Press.

Delgrande, J. P., and Peppas, P. 2015. Belief revision in Horn theories. *Artificial Intelligence* 218:1–22.

Dizadji-Bahmani, F., and Bradley, S. 2014. Lewis' account of counterfactuals is incongruent with Lewis' account of laws of nature. available at http://philsci-archive.pitt.edu/10875/.

Eiter, T., and Gottlob, G. 1992. On the complexity of propositional knowledge base revision, updates, and counterfactuals. *Artif. Intell.* 57(2-3):227–270.

Floridi, L. 2010. Information, possible worlds and the cooptation of scepticism. *Synthese* 175(Suppl 1):63–88.

Friedman, N., and Halpern, J. Y. 1994. On the complexity of conditional logics. In *Principles of Knowledge Representation and Reasoning*, 202–213. Morgan Kaufmann.

Gärdenfors, P. 1984. Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy* 62(2):136–157.

Goldszmidt, M., and Pearl, J. 1992. Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. *KR* 92:661–672.

Grahne, G. 1998. Updates and counterfactuals. *Journal of Logic and Computation* 8(1):87–117.

Grove, A. 1988. Two modellings for theory change. *J. of Philosophical Logic* 17:157–170.

Herzig, A. 1998. Logics for belief base updating. In Dubois, D.; Gabbay, D.; Prade, H.; and Smets, P., eds., *Handbook of defeasible reasoning and uncertainty management*, volume 3 - Belief Change. Kluwer. 189–231.

Huang, X.; Izza, Y.; Ignatiev, A.; Cooper, M.; Asher, N.; and Marques-Silva, J. 2022. Tractable explanations for d-dnnf classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5719–5728.

Katsuno, H., and Mendelzon, A. O. 1991. Propositional knowledge base revision and minimal change. *Artificial Intelligence* 52(3):263–294.

Kenny, E. M., and Keane, M. T. 2021. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*, 11575–11585. AAAI Press.

Kraus, S.; Lehmann, D.; and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and acumulative logics. *Artificial Intelligence* 44(1-2):167–207.

Lewis, D. K. 1973. *Counterfactuals*. Harvard University Press.

Lewis, D. K. 1995. Causation. *Journal of Philosophy* 70(17):556–567.

Liu, X., and Lorini, E. 2021. A logic for binary classifiers and their explanation. In Baroni, P.; Benzmüller, C.; and Wáng, Y. N., eds., *Logic and Argumentation - 4th International Conference, CLAR 2021, Hangzhou, China, 2021, Proceedings*, Lecture Notes in Computer Science, 302–321. Springer.

Liu, X., and Lorini, E. 2023. A unified logical framework for explanations in classifier systems. *Journal of Logic and Computation* 33(2):485–515.

Mittelstadt, B.; Russell, C.; and Wachter, S. 2019. Explaining explanations in AI. In *Proceedings of the 2019 conference on Fairness, Accountability, and Transparency*, 279–288.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.

Pozos-Parra, P.; Liu, W.; and Perrussel, L. 2013. Dalal's revision without hamming distance. In *Advances in Artificial Intelligence and Its Applications: 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part I 12*, 41–53. Springer.

Quine, W. V. O. 1950. *Methods of logic*. Harvard University Press.

Segerberg, K. 1989. Notes on conditional logic. *Studia Logica* 157–168.

Sokol, K., and Flach, P. A. 2019. Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety. In *SafeAI@ AAAI*.

Williamson, T. 1988. First-order logics for comparative similarity. *Notre Dame Journal of Formal Logic* 29(4).