Surprise Minimization Revision Operators

Adrian Haret
Institute for Logic, Language and Computation
The University of Amsterdam
a.haret@uva.nl

Abstract
Prominent approaches to belief revision prescribe the adoption of a new belief that is as close as possible to the prior belief, in a process that, even in the standard case, can be described as attempting to minimize surprise. Here we extend the existing model by proposing a measure of surprise, dubbed relative surprise, in which surprise is computed with respect not just to the prior belief, but also to the broader context provided by the new information, using a measure derived from familiar distance notions between truth-value assignments. We characterize the surprise minimization revision operator thus defined using a set of intuitive rationality postulates in the AGM mould, along the way obtaining representation results for other existing revision operators in the literature, such as the Dalal operator and a recently introduced distance-based min-max operator.

1 Introduction
Belief change models rational adjustments made to an agent’s epistemic state upon acquiring new information (Peppas 2008; Hansson 2017; Fermé and Hansson 2018). When the new information is assumed to be reliable, the logic of changing one’s prior beliefs to accommodate such new-found knowledge falls under the heading of revision. Belief revision is typically thought of as appeal to a set of intuitive normative principles, usually along the lines of the AGM framework (Alchourrón, Gärdenfors, and Makinson 1985), alongside more concrete revision representations and mechanisms (Grove 1988; Dalal 1988; Gärdenfors and Makinson 1988; Katsuno and Mendelzon 1992; Rott 1992).

A perspective underlying many of these representations, which we share here, is that belief revision is akin to a choice procedure guided by a plausibility relation over possible states of affairs: revising a belief, in this sense, amounts to choosing the most plausible states of affairs consistent with the new information. Plausibility over states of affairs, in turn, is judged according to some notion of dissimilarity, or distance between states of affairs: I judge a situation to be less likely the further away from my own belief it is. Among the various distance notions that can be used to make this intuition precise, the approach using Hamming distance to rank truth-value assignments is among the most prominent, used for the well-known Dalal revision operator (Dalal 1988), and the more recently introduced distance-based min-max operator (Haret and Woltran 2019).

Both the Dalal and the Hamming distance min-max operator are designed to respond to new information by minimizing departures from the prior belief, in what can be described, just as well, as an attempt to prevent major surprise: if I have a prior belief that all major carbon emitting countries will have halved their emissions by the end of 2049, and it turns out that neither of them has, then I am likely to be surprised—certainly more surprised than seeing my belief confirmed. Consequently, if I acquire information to the effect that these are the only two possible outcomes (i.e., either all countries cut emissions, or none of them does), then, on the assumption that this information stems from some noisy observation of the true state, I will use my prior belief and gravitate towards the outcome that occasions less surprise.

In this revision procedure, consistent with both the Dalal and the min-max operators, the measure of surprise is taken to depend only on the absolute difference between my prior belief and the states of affairs learned to be viable. However, we can readily imagine that the amount of anticipated surprise depends in equal measure on other factors, e.g., the context provided by the newly acquired information: if in 2049 it turns out that none of the countries has reduced emissions, then I am likely to be less surprised if I had been told in advance that at most one of them would than if I had been told that, possibly, any number of them could achieve the target. In other words, it is desirable to have a broader notion of surprise complementing the absolute one, to account for situations in which change in the epistemic state depends not just on the prior belief but also on the range of options provided by the new information. However, despite the fact that surprise minimization is a natural idea that has been gaining traction in Cognitive Science (Friston 2010; Hohwy 2016), there are not many belief revision policies that explicitly take it into account.

In this paper we put forward a notion of relative surprise that is richer in precisely this sense, and leverage it to define a new type of revision operator, called the Hamming surprise min-max operator, and which is calibrated to take into account contextual effects as described above. Though it deviates from some of the postulates in the AGM framework (notably, Vacuity, Superexpansion and Subexpansion (Fermé and Hansson 2018)), we show that the Hamming surprise min-max operator shares other desirable, though less
obvious, features with the Dalal and the Hamming distance min-max operator. Significantly, we use these features to fully characterize the newly introduced surprise operator, in the process obtaining full characterizations for the Dalal and Hamming distance min-max operators.

Contributions. On a conceptual level, we argue that the notion of distance standardly used to define revision operators can be seen as quantifying a measure of surprise, with different distance-based operators providing different ways to minimize it. We then enrich this landscape by introducing a notion of relative surprise, which is then put to use in defining the Hamming surprise min-max operator. We compare this operator against the standard KM postulates for revision (Katsuno and Mendelzon 1992) and present new postulates that complement the KM ones, for a full characterization. The versatility of the ideas underlying these postulates is showcased by adapting them to the Dalal and Hamming distance min-max operators: in the case of the min-max operator our postulates complement the subset of KM postulates the operator is known to satisfy; in the case of the Dalal operator our postulates strengthen the KM postulates. In both cases, we obtain full characterizations.

Related work. Among belief revision operators that are insensitive to syntax, the Dalal operator has received a significant amount of attention, either from attempts to express it by encoding the Hamming distance between truth-value assignments at the syntactic level (del Val 1993; Pozos-Parra, Liu, and Perrussel 2013); as an instance of the more general class of parameterized difference operators (Peppas and Williams 2018; Aravanis, Peppas, and Williams 2021); or in relation to Parikh’s relevance-sensitivity axiom (Peppas et al. 2015). However, to the best of our knowledge, the characterization we offer here is the first of its kind.

Strengthening the AGM framework to induce additional desired behavior from revision operators has been considered in relation to issues of iterated revision (Darwiche and Pearl 1997), or relevance sensitivity (Parikh 1999; Peppas and Williams 2016). In terms of choice rules, the closest analogue to the surprise minimization operator is the decision rule that minimizes maximum regret in decisions with ignorance (Milnor 1954; Lave and March 1993; Peterson 2017), with Hamming distances playing the role of utilities in our present setting. However, the logical setting and the fact that the distances depend on the states themselves means that decision theoretic results do not translate easily to our current framework.

Outline. Section 2 introduces the main notions related to propositional logic and belief revision that will be used in the rest of the paper, and argues for the surprise-based interpretations of distances, Section 3 defines the relative Hamming surprise measure and the Hamming surprise min-max operator. Sections 4 and 5 consist of a slight detour in which the Dalal and Hamming distance min-max operators are characterized, setting up the stage for the characterization of the surprise operator in Section 6. Section 7 offers conclusions.

2 Preliminaries

Propositional Logic. We assume a finite set $A$ of propositional atoms, large enough that we can always reach into it and find additional, unused atoms, if any are needed. The set $L$ of propositional formulas is generated from the atoms in $A$ using the usual propositional connectives ($\land$, $\lor$, $\neg$, $\rightarrow$ and $\leftrightarrow$), as well as the constants $\bot$ and $\top$.

An interpretation $w$ is a function mapping every atom in $A$ to either true or false. Since an interpretation $w$ is completely determined by the set of atoms in $A$ it makes true, we will identify $w$ with this set of atoms and, if there is no danger of ambiguity, display $w$ as a word where the letters are the atoms assigned to true. The universe $U$ is the set of all interpretations for formulas in $L$. If $w_1$ and $w_2$ are interpretations, the symmetric difference $w_1 \triangle w_2$ of $w_1$ and $w_2$ is defined as $w_1 \triangle w_2 = (w_1 \setminus w_2) \cup (w_2 \setminus w_1)$, i.e., as the set of atoms on which $w_1$ and $w_2$ differ. The Hamming distance $d_H: U \times U \rightarrow \mathbb{N}$ is defined, for any interpretations $w_1$ and $w_2$, as $d_H(w_1, w_2) = |w_1 \triangle w_2|$. Intuitively, the Hamming distance $d_H(w_1, w_2)$ between $w_1$ and $w_2$ counts the number of atoms that $w_1$ and $w_2$ differ on, and is used to quantify the disagreement between two interpretations.

The models of a propositional formula $\phi$ are the interpretations that satisfy it, and we write $[\phi]$ for the set of models of $\phi$. If $\phi_1$ and $\phi_2$ are propositional formulas, we say that $\phi_1$ entails $\phi_2$, written $\phi_1 \models \phi_2$, if $[\phi_1] \subseteq [\phi_2]$, and that they are equivalent, written $\phi_1 \equiv \phi_2$, if $[\phi_1] = [\phi_2]$. A propositional formula $\phi$ is consistent if $[\phi] \neq \emptyset$. The models of $\bot$ and $\top$ are $[\bot] = \emptyset$ and $[\top] = U$. We will occasionally find it useful to explicitly represent the models of a formula, in which case we write $\{\phi_1, \ldots, \phi_n\}$ for a propositional formula such that $\bigcap_{i=1}^n [\phi_i] = \emptyset$. A propositional formula $\phi$ is complete if it has exactly one model, and we will typically denote a complete formula as $\phi_\ast$, to draw attention to its unique model $\ast$. The null formula $\varepsilon$ and the full formula $\alpha$ are defined as $\varepsilon = \bigwedge_{p \in A} \neg p$ and $\alpha = \bigwedge_{p \in A} p$, i.e., as the conjunction of the negated and non-negated atoms in $A$, respectively. Note that $[\varepsilon] = \{\emptyset\}$ and $[\alpha] = A$.

Distance-based belief revision. A revision operator $\circ$ is a function $\circ: L \times L \rightarrow L$, taking as input two propositional formulas, denoted $\phi$ and $\mu$, and standing for the agent’s prior and newly acquired information, respectively, and returning a propositional formula, denoted $\phi \circ \mu$. Two revision operators $\circ_1$ and $\circ_2$ are equivalent, written $\circ_1 \equiv \circ_2$, if $\phi \circ_1 \mu \equiv \phi \circ_2 \mu$, for any formulas $\phi$ and $\mu$.

The primary device for generating concrete revision operators we make recourse to here is the Hamming distance. Thus, the Hamming distance min-min operator $\circ_{d_H, \min}$, or, as it is more commonly known, the Dalal operator (Dalal 1988), is defined, for any propositional formulas $\phi$ and $\mu$, as a formula $\phi \circ_{d_H, \min} \mu$ such that:

$$[\phi \circ_{d_H, \min} \mu] = \arg\min_{w \in [\mu]} \min_{v \in [\phi]} d_H(v, w).$$

Intuitively, the shortest distance from $w$ to any model of $\phi$, i.e., $\min_{v \in [\phi]} d_H(v, w)$, can be interpreted as a measure of distance between $w$ and $\phi$, and we will refer to it as the Hamming min-distance between $\phi$ and $\mu$. The result $\phi \circ_{d_H, \min} \mu$
of revision, then, selects those models of $\mu$ that are closest to $\phi$ according to this measure.

Recently, an alternative revision operator has been analyzed (Haret and Woltran 2019): what we will call here the \textit{Hamming distance min-max operator} $\phi \circ \text{d}_H^{\text{min}}, \text{max}$, defined, for any $\phi$ and $\mu$, as a formula $\phi \circ \text{d}_H^{\text{min}}, \text{max} \mu$ such that:

$$ [\phi \circ \text{d}_H^{\text{min}}, \text{max} \mu] = \arg\min_{w \in [\mu]} \max_{v \in [\phi]} d_H(v, w),$$

i.e., a formula whose models are exactly those models of $[\mu]$ that minimize the Hamming distance to $\text{max}_{v \in [\phi]} d_H(v, w)$, the Hamming max-distance between $\phi$ and $\mu$.

**Distance as surprise.** Consistent with the idea that revision models the agent learning about the world around it, we can see the new information $\mu$ as a noisy observation of some underlying ground truth state $w^*$: by acquiring $\mu$, the agent learns of a set of outcomes (the models of $\mu$), all of which stand a chance of being the true state $w^*$. In that sense, the distance $d(v, w)$ between any $v \in [\phi]$ and $w \in [\mu]$ stands for a quantity that can be aptly described as \textit{surprise}: it is the difference between what the agent expects is the case ($v$) and what might turn out to actually be the case ($w$). Naturally, the agent will want to minimize the divergence between its predictions and reality, with existing revision operators providing different means to do so.

**Example 1.** Consider a set $A = \{a, b, c, d, e\}$ of atoms, standing for countries that might meet their emission targets before 2049, and formulas $\phi = (a \land \neg b \land \neg c \land \neg d \land \neg e) \lor (a \land b \land c \land d \land e)$ and $\mu = \phi \lor (a \land b \land \neg c \land \neg d \land e)$, with $[\phi] = \{\emptyset, abcd\}$ and $[\mu] = \{\emptyset, abcd, a\}$. Using the Hamming distances depicted in Table 1, we obtain that $[\phi \circ \text{d}_H^{\text{min}} \mu] = \{\emptyset, abcd\}$ and $[\phi \circ \text{d}_H^{\text{max}} \mu] = \{abc\}$. Intuitively, we read this as saying that if an agent believes the true state to be either of the worlds in $[\phi] = \{\emptyset, abcd\}$, but finds out it is one among $[\mu] = \{\emptyset, abcd, a\}$, then $\circ \text{d}_H^{\text{min}}$ selects the new belief to be $\{\emptyset, abcd\}$, as this supplies the least amount of surprise in an optimistic, best best-case scenario: if the true state turns out to be either of $\emptyset$ or $abcd$, then the agent, believing this, will be able to say “I told you so!”; the $\text{abcd}$ case, which is surprising in both cases, is ignored. In a complementary approach, the $\circ \text{d}_H^{\text{max}}$ operator shifts the agent’s belief to $\{abcd\}$, as this provides, more cautiously, the best worst-case scenario:

| $d_H$ | $\emptyset$ | abcd | min | max |
|------|----------|------|-----|-----|
| $\emptyset$ | 0 | 4 | 0 | 4 |
| abcd | 4 | 0 | 0 | 4 |
| $abe$ | 3 | 3 | 3 | 3 |

Table 1: Hamming distances $d_H(v, w)$ for $v \in [\phi]$, $w \in [\mu]$, with $[\phi] = \{\emptyset, abcd\}$ and $[\mu] = \{\emptyset, abcd, a\}$. The lower and maximal values per model of $\mu$ are tallied on the right, with the values preferred by operators $\circ \text{d}_H^{\text{min}}$ and $\circ \text{d}_H^{\text{max}}$, i.e., the minimal among the minimal and maximal values, respectively, in bold font.

from the standpoint of both $\emptyset$ or abcd, $abe$ seems the least risky of the other options.

Example 1 serves as a springboard for some important observations. Firstly, it illustrates that $\circ \text{d}_H^{\text{min}}$ and $\circ \text{d}_H^{\text{max}}$ are distinct operators. Secondly, it is apparent from Example 1 that, given prior beliefs $\phi$, interpretations can be ranked according to their Hamming min- or max-distance to $\phi$. It is straightforward to see that (i) in both cases the resulting rankings depend only on the models of $\phi$, are total and admit ties; (ii) the min-distance places models of $\phi$ at the bottom of this ranking, i.e., as the most plausible interpretations according to $\phi$, in a pattern that goes under the name of a faithful ranking (Katsuno and Mendelzon 1992); and, perhaps, less conspicuously, that (iii) the max-distance places models of the so-called dual of $\phi$ (i.e., the formula obtained from $\phi$ by replacing all its atoms with their negations), at the very top, i.e., as the least plausible interpretations according to $\phi$ (Haret and Woltran 2019). The different flavors of rankings, faithful or otherwise, generated in this distance-based approach usually play a prominent role in representation results for revision, as they open up a level of abstraction between that of concrete numbers and general principles. In this work, however, we will bypass talk of rankings and work directly at the interface between distance-based measures and normative principles.

Finally, an observation that will prove useful is that we can (and will) think of the individual models $v$ of $\phi$ as generating their own plausibility rankings over interpretations: these rankings correspond to the columns in Table 1 and are the rankings that would be generated if the prior belief were the complete formula $[\phi] = \{v\}$, i.e., what the landscape of plausibility looks like if the agent puts the entire weight of its belief on $\phi$. Revision can then be seen as employing a function (min or max) to aggregate the individual rankings, and then choosing something out of the aggregated result: the Dalal operator $\circ \text{d}_H^{\text{min}}$ chooses, optimistically, the models that are the best of the best, while $\circ \text{d}_H^{\text{max}}$ chooses, pessimistically, the best of the worst models across the individual rankings. In keeping with this way of looking at things, we will often speak, loosely, of formulas and interpretations ‘judging’ and ‘choosing’ among possible outcomes.

What recommends the choice behavior of operators (such as Dalal’s operator) as reasonable is adherence to a set of intuitive normative principles, or rationality postulates. The most common set of such principles consists of the AGM postulates for revision (Alchourrón, Gärdenfors, and Makinson 1985), which we present here in the Katsuno-Mendelzon formulation (Katsuno and Mendelzon 1992). The postulates apply for any propositional formulas $\phi$, $\mu$, $\mu_1$ and $\mu_2$:

(R1) $\phi \circ \mu \models \mu$.

(R2) If $\phi \land \mu$ is consistent, then $\phi \circ \mu \equiv \phi \land \mu$.

(R3) If $\mu$ is consistent, then $\phi \circ \mu$ is consistent.

(R4) If $\phi_1 \equiv \phi_2$ and $\mu_1 \equiv \mu_2$, then $\phi_1 \circ \mu_1 \equiv \phi_2 \circ \mu_2$.

(R5) $(\phi \circ \mu_1) \land \mu_2 \models \phi \circ (\mu_1 \land \mu_2)$.

(R6) If $(\phi \circ \mu_1) \land \mu_2$ is consistent, then $\phi \circ (\mu_1 \land \mu_2) \models (\phi \circ \mu_1) \land \mu_2$. 

The primary assumption of revision (postulate R₁) is that new information originates with a trustworthy source; thus, revising \( \varphi \) by \( \mu \) involves a commitment to accept the newly acquired information. Postulate R₂, known as the Vacuity postulate, says that if the newly acquired information \( \mu \) does not contradict the prior information \( \varphi \), the result is just the conjunction of \( \mu \) and \( \varphi \). Postulate R₃ says that if the newly acquired information \( \mu \) is consistent, then the revision result should also be consistent. Postulate R₄ says that the result depends only on the semantic content of the information involved. Postulates R₅ and R₆, known as Subexpansion and Superexpansion, respectively, ensure a certain kind of coherence when the new information is presented sequentially, which is for present purposes best understood as akin to a form of independence of irrelevant alternatives familiar from rational choice (Sen 1977): the choice over two alternatives (here, interpretations \( w₁ \) and \( w₂ \) in \( [\mu] \)) should not depend on the presence of other alternatives in the menu (here represented by new information \( \mu \)).

The Dalal operator \( \phi_{\text{h}_\mu} \) satisfies postulates R₁-R₆ (Katsuno and Mendelzon 1992), though these postulates do not uniquely characterize it. The Hamming distance max-operator \( \phi_{\text{h}_\mu}^{\max} \) satisfies postulates R₁ and R₃-R₆ but not R₂, though it does satisfy the following two postulates (Haret and Woltran 2019), where \( \varphi \) stands for the dual of \( \varphi \), as defined above:

\[(R₂) \text{ If } \varphi \circ \mu \models \varphi, \text{ then } \varphi \circ \mu \equiv \mu.\]

\[(R₆) \text{ If } \mu \not\models \varphi, \text{ then } (\varphi \circ \mu) \land \varphi \text{ is inconsistent.}\]

In certain circumstances, \( \varphi \) can be thought of as the point of view opposite to that of \( \varphi \), such that, taken together, postulates R₇ and R₈ inform the agent to believe states of affairs compatible with \( \varphi \) only if it has no other choice in the matter: the models of \( \varphi \) should be part of a viewpoint one is willing to accept only as a last resort.

### 3 Relative Hamming Surprise Minimization

In this section we introduce our novel surprise-based operator. We start by defining, for any interpretations \( v \) and \( w \), the (relative) Hamming surprise \( s^\mu_{\text{h}}(v, w) \) of \( v \) with respect to \( w \) relative to \( \mu \), as:

\[s^\mu_{\text{h}}(v, w) = d_\text{h}(v, w) - d_\text{h}(v, \mu),\]

i.e., the distance between \( v \) and \( w \) normalized by the distance between \( v \) and \( \mu \). The new information \( \mu \), here, serves as the reference point, or context, relative to which surprise is calculated. The Hamming surprise min-max operator \( \phi_{\text{h}^{\min}, \max} \) is defined as a formula \( \varphi \circ_{\text{h}^{\min}, \max} \mu \) such that:

\[|\varphi \circ_{\text{h}^{\min}, \max} \mu| = \arg\min_{\mu \in [\mu]} \max_{\varphi \in [\varphi]} s^\mu_{\text{h}}(v, w),\]

i.e., as a formula whose models are exactly those models of \( \mu \) that minimize maximum Hamming surprise with respect to \( \varphi \), and relative to \( \mu \). We refer to \( \max_{\varphi \in [\varphi]} s^\mu_{\text{h}}(v, w) \), as the max-surprise of \( \varphi \) with \( w \) relative to \( \mu \).

**Example 2.** Consider formulas \( \varphi \) and \( \mu \) as in Example 1, with \( [\varphi] = \{\emptyset, \text{abcd}\} \) and \( [\mu] = \{\emptyset, \text{abcd}, \text{abe}\} \). We have that \( d_\text{h}(\emptyset, \mu) = \min_{\mu \in [\mu]} d_\text{h}(\emptyset, w) = 0 \), and thus

| \( s^\mu_{\text{h}} \) | \( \emptyset \) | \( \text{abcd} \) | \( \text{max} \) |
|---|---|---|---|
| \( \emptyset \) | 0 | 0 | 0 |
| \( \text{abcd} \) | 4 | 0 | 0 |
| \( \text{abe} \) | 3 | 3 | 0 |

Table 2: Relative Hamming surprise \( s^\mu_{\text{h}}(v, w) \) for \( v \in [\varphi], \ w \in [\mu], \) for \( [\varphi] = \{\emptyset, \text{abcd}\}, \ [\mu] = \{\emptyset, \text{abcd}, \text{abe}\}, \) and relative to \( \mu: \ d_\text{h}(v, w) \) is normalized by the distance \( d_\text{h}(v, \mu) \) from \( v \) to \( \mu \). The lower surprise is, the more plausible \( w \) is considered to be, from the standpoint of \( v \). The model minimizing overall surprise is emphasized in bold font.

| \( s^\mu_{\text{H}} \) | \( \emptyset \) | \( \text{abcd} \) | \( \text{max} \) |
|---|---|---|---|
| \( \emptyset \) | 4 | 3 | 0 |
| \( \text{abe} \) | 3 | 3 | 0 |

Table 3: Relative Hamming surprise \( s^\mu_{\text{H}}(v, w) \), \( [\varphi] = \{\emptyset, \text{abcd}\}, \ [\mu] = \{\emptyset, \text{abcd}, \text{abe}\} \). The best interpretation is now \( \text{abcd} \): the ranking induced by relative surprise depends on \( \mu \), as well as \( \varphi \).
the new information, as in Example 2, can lead to inversions between the relative ranking of two interpretations. At the same time, the ranking plainly depends on nothing more than \( \varphi \) and \( \mu \), such that the result of revision is invariant to the syntax of the prior and new information. Additionally, \( \circ^{\#H, \max} \) selects the result from the models of \( \mu \), and is guaranteed to output something as long as \( \mu \) is consistent. We summarize these observations in the following proposition.

**Proposition 1.** The operator \( \circ^{\#H, \max} \) satisfies postulates R1, R3 and R4, but not R2, R5, R6, R7 and R8.

One detail worth mentioning is that when \( \varphi \) is complete all operators presented so far coincide.

**Proposition 2.** For any complete formula \( \varphi \), \( \varphi \circ^{\#H, \min} \mu \equiv \varphi \circ^{\#H, \max} \mu \), for any formula \( \mu \).

**Proof.** For complete \( \varphi \), it is only the relative ranking of interpretations with respect to \( \varphi \) that matters, and this is the same for all three operators. \( \square \)

Proposition 1 shows that the \( \circ^{\#H, \max} \) operator does not fit neatly into the standard revision framework. However, since, we have argued, \( \circ^{\#H, \max} \) formalizes an appealing intuition, it will be useful to unearth the general rules underpinning it: our goal, now, is to find a set of normative principles strong enough to characterize \( \circ^{\#H, \max} \). A set of such principles is offered in Section 6, but, since \( \circ^{\#H, \max} \) can be seen as a more involved min-max operator, we set the scene by first characterizing \( \circ^{\#H, \min} \). And to set the scene for \( \circ^{\#H, \max} \), we first characterize the Dalal operator.

### 4 Characterizing the Dalal Operator

In this section we present a set of postulates that characterize the Dalal operator \( \circ^{\#H, \min} \). Apart from being of independent interest, this section presents, in the familiar setting of a known operator, the main intuitions and techniques used in subsequent sections. We start by introducing some additional new notions.

A *renaming* \( r \) of \( A \) is a bijective function \( r : A \rightarrow A \). If \( \varphi \) is a propositional formula, the renaming \( r(\varphi) \) of \( \varphi \) is a formula \( r(\varphi) \) whose atoms are replaced according to \( r \). On the semantic side, if \( w \) is an interpretation and \( r \) is a renaming of \( A \), the renaming \( r(w) \) of \( w \) is an interpretation obtained by replacing every atom \( p \) in \( w \) with \( r(p) \). If \( W \) is a set of interpretations, the renaming \( r(W) \) of \( W \) is defined as \( r(W) = \{ r(w) \mid w \in W \} \), i.e., the set of interpretations whose elements are the renamed interpretations in \( W \).

A *flip function* \( f : 2^A \times L \rightarrow L \) is a function that takes as input a set \( v \subseteq A \) of atoms (equivalently, \( v \) can be thought of as an interpretation) and a propositional formula \( \varphi \), and returns a propositional formula \( f_v(\varphi) \) that is just like \( \varphi \) except that all the atoms from \( v \) that appear in \( \varphi \) are flipped, i.e., replaced with their negations. Overloading notation, a flip function applied to interpretations \( v \) and \( w \) returns an interpretation \( f_v(w) \) in which all the atoms from \( v \) that appear in \( w \) are flipped, i.e., \( f_v(w) = \{ p \in A \mid p \in w \land p \notin v, \text{ or } p \in v \land p \notin w \} \). It is straightforward to see that \( f_v(w) = w \triangle v \). If \( W \) is a set of interpretations, then \( f_v(W) = \{ f_v(w) \mid w \in W \} \), i.e., the set of interpretations obtained by flipping every atom in \( v \).

**Example 3.** For the set \( A = \{a, b, c\} \) of atoms, consider a formula \( \varphi = a \land \neg c \), with \( \{\varphi\} = \{a, \beta b\} \), and a renaming \( r \) such that \( r(a) = b, r(b) = c \) and \( r(c) = a \). We obtain that \( r(\varphi) = (a \land \neg c) = b \land \neg a \). Flipping atoms \( b \) and \( c \), we have that \( f_{bc}(\varphi) = a \land \neg c \land b \land \neg a \). Flipping atoms \( b \) and \( c \), we have that \( f_{bc}(\varphi) = a \land \neg c \land b \land \neg a \).

In Example 3 it holds that: (i) \( [r(\varphi)] = r([\varphi]) \), (ii) \( [f_w(\varphi)] = f_w([\varphi]) \) and (iii) \( [f_{bc}(\varphi)] = \{a \land \neg c \land b \land \neg a \} \).

Note that \( [f_{bc}(\varphi)] = \{a \land \neg c, ab \land \neg be \} \).

Postulate \( R_4 \) states that revision is invariant under renaming atoms and hence neutral in that the specific labels for the atoms do not matter towards the final result. This postulate is inspired by similar ideas in social choice and has appeared before in belief change contexts (Herzig and Rifi 1999; Marquis and Schwinding 2014; Haret and Woltran 2019).

The next postulate concerns the effect of flipping the same atoms in both \( \varphi \) and \( \mu \), and is called, appropriately, the *Flipping* postulate \( R_5 \).

We obtain that \( r(\varphi \circ \mu) = r(\varphi) \circ r(\mu) \).

An additional constraint, the *Addition* postulate \( R_4 \), is obtained by considering the effect of adding new atoms that affect the standing of one interpretation, and is meant to apply to any formulas \( \varphi \) and \( \mu \) and set \( x \) of new atoms, i.e., such that none of the atoms in \( x \) appears in either \( \varphi \) or \( \mu \): (R4) If \( \varphi \) is complete and \( (\varphi \circ \mu_{w_1, w_2}) \land \mu_{w_1} \) is consistent, then \( \varphi \circ \mu_{w_1, w_2} \cup x \land \mu_{w_1} \) is consistent.

Postulate \( R_4 \) is best understood through a choice perspective: if \( w_1 \) is chosen by \( \varphi \) over \( w_2 \) when the choice is \( \{\mu_{w_1, w_2}\} = \{w_1, w_2\} \), then adding extra new atoms \( x \) to \( w_2 \), and, thereby, increasing the distance to \( \varphi \) ensures that \( w_2 \cup x \) is not chosen when the choice is \( \{\mu_{w_1, w_2} \cup x\} = \{w_1, w_2 \cup x\} \). In all of these postulates the prior belief \( \varphi \) is assumed to be complete: this is not essential for the characterization of the Dalal operator, but makes life easier in the characterization of the surprise minimization operator, in Section 6.

The next postulate involves a mix of flips and we ease into it by introducing an intermediary notion. The *best-of-best* formula \( \beta_{\varphi, \mu} \) with respect to \( \varphi \) and \( \mu \) is defined as:

\[
\beta_{\varphi, \mu} = \epsilon \circ \left( \bigvee_{v \in [\varphi]} f_v(\mu) \right),
\]

i.e., as the result of revising the null formula \( \epsilon \) (recall that \( [\epsilon] = \{\emptyset\} \) by a disjunction made up of multiple versions of \( \mu \), where each such version is obtained by flipping the atoms in a model \( v \) of \( \varphi \). Intuitively, the intention is to recreate the table of Hamming distances (e.g., Table 1) without using numbers: recall that \( f_v(\mu) = f_v(\mu) \) and \( f_v(w) = w \triangle v \) and thus, semantically, we have that \( \bigvee_{v \in [\varphi]} f_v(\mu) = \epsilon \circ \left( \bigvee_{v \in [\varphi]} f_v(\mu) \right) \).
Applying the Neutrality postulate $R_{\text{BOB}}$, then, is to recover the models of $\mu$ from the models of $\beta_{\varphi,\mu}$:

$$(R_{\text{BOB}}) \quad \varphi \circ \mu \equiv \left( \bigvee_{v \in \varphi} f_{\circ}(\beta_{\varphi,\mu}) \right) \wedge \mu.$$ 

Postulate $R_{\text{BOB}}$ stipulates that the result of revising $\varphi$ by $\mu$ consists of those interpretations of $\mu$ that come out of flipping $\beta_{\varphi,\mu}$, by each model of $\varphi$, in this way reversing the initial flips that delivered the revision formula posed to $\varepsilon$.

What is the significance of the null formula $\varepsilon$ in $\beta_{\varphi,\mu}$? We want to reduce arbitrary revision tasks to a common denominator, a base case in which the result of revision can be decided without explicit appeal to distances (i.e., numbers), and only by appeal to desirable normative principles, such as the postulates laid out above. The case when the prior belief is $\varepsilon$ turns out to be well suited for this task, since, as we show next, postulates $R_1$, $R_3-R_6$, $R_8$, and $R_9$ guarantee that $\varepsilon$ always selects the interpretations with minimal cardinality.

**Lemma 1.** If a revision operator $\circ$ satisfies postulates $R_1$, $R_3-R_6$, $R_8$ and $R_9$, then, for any formula $\mu$, it holds that $[\varepsilon \circ \mu] = \arg\min_{w \in [\mu]} |w|$.

**Proof.** ("$\subseteq\)"") Suppose, first, that $w_1 \in [\varepsilon \circ \mu]$ and there is $w_2 \in [\mu]$ such that $w_1 > w_2$. Using postulate $R_5$ we obtain that $w_1 \in [\varphi \circ \mu]$, $w_2$. We now show that this leads to a contradiction, and we do this using the Neutrality postulate $R_8$: however, we would like to apply $R_8$ to interpretations of equal size. Towards this, take a set $x$ of new atoms (i.e., that do not occur in either $\varphi$ or $\mu$), with $|x| = |w_1 - w_2|$, and add $x$ to $w_1$ to form $w_1 \triangle v = w_2 \cup x$. We have that $w_1 \triangle v = w_2 \triangle v = w_1 \triangle w_2$, i.e., $w_1$ and $w_2$ are of the same size, which implies that $w_1 \backslash w_2 = w_2 \backslash w_1$. Applying the addition postulate $R_8$, we obtain that $w_2 \notin [\varepsilon \circ \mu]_{w_1 \triangle v}$.

Consider, now, a renaming $r$ that swaps atoms in $w_1 \backslash w_2$ with atoms in $w_1 \backslash w_2$, made possible by the fact that $w_1 \backslash w_2$ and $w_2 \backslash w_1$ are of the same size. This implies that $r(w_1) = w_2$ and $r(w_2) = w_1$ and thus $r([\mu_{w_1,w_2}]) = r([w_1,w_2]) = [r(w_1), r(w_2)] = [w_2, w_1] = [\mu_{w_1,w_2}]$. Applying the Neutrality postulate $R_8$ to $\varepsilon \circ \mu_{w_1,w_2}$ with the renaming $r$ thus defined, and, keeping in mind that $[r(\varepsilon)] = [\varepsilon]$, and thus that $r(\varepsilon) \equiv \varepsilon$, we obtain that:

$$\begin{align*}
\{w_1\} &= [\varepsilon \circ \mu_{w_1,w_2}] \\
= [r(\varepsilon) \circ r([\mu_{w_1,w_2}])] &= \text{by def. of } r \text{ and } R_4 \\
= [r(\varepsilon \circ \mu_{w_1,w_2})] &= \text{by } R_4 \\
= r([\varepsilon \circ \mu_{w_1,w_2}]) &= \text{property of } r \\
= \{w_1\} &= \text{by assumption}.
\end{align*}$$

This implies that $w_1 = w_2$, but, since $w_2$ contains a non-negative number of atoms that do not appear in $w_1$, this is a contradiction.

("$\supseteq\)"") For the opposite direction, suppose that $w_1 \in \arg\min_{w \in [\mu]} |w|$ but $w_1 \notin [\varepsilon \circ \mu]$. Using postulates $R_1$ and $R_5$ we have that there is $w_2 \in [\varphi \circ \mu]$ and, with postulate $R_6$, we obtain that $[\varepsilon \circ \mu_{w_1,w_2}] = [w_2]$. Since $w_1 \leq w_2$ we add to $w_1$ a set $x$ of new atoms, where $|x| = |w_2 - w_1|$, and denote $w'_1 = w_1 \cup x$. Applying $R_8$ we obtain that $[\varepsilon \circ \mu_{w'_1,w_2}] = [w_2]$ and, using a renaming $r$ defined, as in the previous direction, such that $r(w_2) = w'_1$ and $r(w'_1) = w_2$, and applying $R_9$ to $r$ and $\varepsilon \circ \mu_{w'_1,w_2}$, we obtain that $[\varepsilon \circ \mu_{w'_1,w_2}] = \{w'_1\}$, leading to a contradiction.

Lemma 1 shows that, in the very particular case in which the prior belief is $\varepsilon$, we can ensure that the result of revision coincides with the result delivered by the Dalal operator. The next move consists in using the Flipping postulate $R_{\varepsilon}$ to extend this fact to complete formulas.

**Lemma 2.** If a revision operator $\circ$ satisfies postulates $R_1$, $R_3-R_6$, $R_8$, $R_9$ and $R_{\varepsilon}$, then, for any formula $\mu$ and complete formula $\varphi_\mu$, it holds that $[\varphi_\mu \circ \mu] = \arg\min_{w = [\mu]} d_H(w, v)$.

**Proof.** By postulate $R_{\varepsilon}$ it holds that $f_{\varepsilon}(\varphi_\mu \circ \mu) \equiv f_{\varepsilon}(\varphi_\mu \circ \mu)$. Note, now, that $[f_{\varepsilon}(\varphi_\mu)] = [v \triangle v] = \emptyset$, and thus $f_{\varepsilon}(\varphi_\mu) \equiv \varepsilon$, while $[f_{\varepsilon}(\mu)] = [v \triangle v] \wedge |w|$. By Lemma 1, it holds that $[\varepsilon \circ [f_{\varepsilon}(\mu)]] = \min_{v \in \Delta v = [\beta_{\varphi,\mu}]} |w\Delta v|$, and, since $d_H(v, w) = |w\Delta v|$, we derive the conclusion.

Lemma 2 shows that it is not just the formula $\varepsilon$ that makes choices consistent with the Dalal operator, but any complete formula $\varphi_\mu$. The intuition driving Lemma 2 is that the situation where $v$ chooses between $w_1$ and $w_2$ is equivalent, through the Flipping postulate $R_{\varepsilon}$ to a scenario where $v$ chooses between $w_1 \Delta v$ and $w_2 \Delta v$; and we know that in this situation postulates $R_8$ and $R_9$ guide $\emptyset$ to choose the interpretation $w \Delta v$ of minimal cardinality, which corresponds to $w_1$ being at minimal Hamming distance to $v$.

The next step involves pushing this intuition even further, to the case of any propositional formula $\varphi$. As anticipated, the Best-of-Best postulate $R_{\text{BOB}}$ is the postulate that facilitates this move, and the proof goes through the intermediary observation that the best-of-best formula $\beta_{\varphi,\mu}$ selects interpretations corresponding to the desired reduced.

**Lemma 3.** If $\circ$ is a revision operator that satisfies postulates $R_1$, $R_3-R_6$, $R_8$, $R_9$ and $R_{\varepsilon}$, then, for any formulas $\varphi$ and $\mu$ and interpretations $w$ and $v$, it holds that $w \Delta v \in [\beta_{\varphi,\mu}]$ if and only if $w \in \arg\min_{w \in [\mu]} \min_{v \in [\varphi]} d_H(w, v)$.

**Proof.** By Lemma 1, $[\beta_{\varphi,\mu}]$ chooses exactly those interpretations $w_i \Delta v_j$, for $w_i \in [\mu]$ and $v_j \in [\varphi]$, that are of minimal cardinality. Since $|w_i \Delta v_j| = d_H(w_i, v_j)$, the conclusion follows immediately.
models of $\mu$ that should be chosen by an operator $\odot$, and provides the final piece in the sought after characterization.

**Theorem 1.** A revision operator $\odot$ satisfies postulates $R_1$, $R_3$, $R_6$, $R_b$, $R_T$ and $R_{\text{BGB}}$ if and only if $\odot = \odot_{\text{H}, \text{min}}$.

**Proof.** For one direction, we take as known that the Dalal operator $\odot_{\text{H}, \text{min}}$ satisfies postulates $R_1$, $R_3$, $R_6$ (Katsuno and Mendelzon 1992) and $R_b$ (Haret and Woltran 2019). For postulate $R_T$, given Lemma 3, satisfaction of postulates $R_b$, $R_6$, $R_T$ and $R_{\text{BGB}}$ follows straightforwardly.

For the other direction, we have to show that if $\odot$ satisfies all the stated postulates, then $[\varphi \odot \mu] = \arg\min_{w \in [\mu]} \min_{v \in [\varphi]} d_H(v, w)$, for any formulas $\varphi$ and $\mu$. Lemma 3 already gives us that $\beta_{\varphi, \mu}$ selects those interpretations $w_i \triangle v_j$ for which $d_H(w_i, v_j)$ is minimal among the set of symmetric differences between models of $\varphi$ and $\mu$. This means that if $w_i \in \arg\min_{w \in [\mu]} \min_{v \in [\varphi]} d_H(v, w)$, then $w_i \triangle v_j \in \beta_{\varphi, \mu}$, for some $v_j \in [\varphi]$, and hence $(w_i \triangle v_j) \triangle v_j = w_i \in [f_v(\beta_{\varphi, \mu})] \triangle v_j$, i.e., if $w_i$ is selected by the Dalal operator then it shows up in $(\bigvee_{v \in [\varphi]} f_v(\beta_{\varphi, \mu})) \triangle v_j$. Conversely, suppose there is an interpretation $w_i \in (\bigvee_{v \in [\varphi]} f_v(\beta_{\varphi, \mu})) \triangle v_j$, and thus $(w_i \triangle v_j) \triangle v_j = w_i \in [\beta_{\varphi, \mu}]$, which is that at minimal distance to $\varphi$. This means that $w_i = (w_i \triangle v_j) \triangle v_j$, where $w_i \in [\mu]$ corresponds to a model of $\mu$ that is at minimal Hamming distance to $\varphi$ and $v_k$, $v_j \in [\varphi]$. We infer from this that $w_i \triangle v_j = (w_i \triangle v_k) \triangle v_j = w_i \triangle v_k$, and thus $|w_i \triangle v_j| = |w_i \triangle v_k|$. But this contradicts the assumed minimality of $w_i \triangle v_k$. \hfill $\square$

Note that postulate $R_2$ is not present in Theorem 1, even though the Dalal operator satisfies it, as it follows from the other postulates.

Theorem 1 can be read not just as a characterization of the Dalal operator, but also as a recipe, or a step-by-step argument, for constructing $\odot_{\text{H}, \text{min}}$ from a set of simpler problems, in a sequence of steps guided by the transformations inherent in postulates $R_b$, $R_6$, $R_T$ and $R_{\text{BGB}}$. The form such an argument could take is illustrated in the following example.

**Example 4.** Consider formulas $[\varphi] = \{a, b\}$ and $[\mu] = \{ac, abc\}$ and note, first, that $[\varphi \odot_{\text{H}, \text{min}} \mu] = \{ac\}$, as $ac$ minimizes overall distance to $\varphi$ via $d_H(a, ac) = 1$. Assume, however, that we are given a revision operator $\odot$ that is not defined using distances, but is presented only as satisfying postulates $R_1$, $R_3$, $R_6$, $R_b$, $R_T$ and $R_{\text{BGB}}$. An agent revising according to $\odot$ can use the postulates to work its way toward $[\varphi \odot_{\text{H}, \text{min}} \mu]$ without knowing anything about distances. This can be done by, first, splitting the problem into two revision problems, one for each model of $\varphi$, $\varphi_a \odot \mu$ and $\varphi_b \odot \mu$, where $[\varphi_a] = \{a\}$ and $[\varphi_b] = \{b\}$. The next step consists in reducing both problems to the common denominator of revising with prior belief $\varepsilon$, where $[\varepsilon] = \{\emptyset\}$. This is done by flipping $a$ and $b$, respectively, in the two problems, to obtain the revision scenarios $\varepsilon \odot f_a(\mu)$ and $\varepsilon \odot f_b(\mu)$, with $[f_a(\mu)] = \{f_a(ac), f_a(abc)\} = \{ac \triangle a, abc \triangle a\} = \{ac, bc\}$ and, likewise, $[f_b(\mu)] = \{abc, ac\}$ (see Figure 1). This move preserves Hamming distances in a crucial way: to take one instance, $d_H(a, ac) = 1$, where $a \in [\varphi]$ and $ac \in [\mu]$, coincides with the Hamming distance between $\emptyset$ and $f_a(ac) = c$, and this distance coincides with the number of atoms in $f_a(ac) = c$. The operator $\odot$, of course, knows nothing of this: it performs these transformations solely because postulate $R_{\text{BGB}}$ warrants it. Thus, in the next step $\odot$ chooses among the models obtained from the successive flips of $\mu$, i.e., it solves the revision problem $\varepsilon \odot (f_a(\mu) \lor f_b(\mu))$. Postulates $R_1$, $R_3$, $R_6$, $R_b$ and $R_T$, via the argument in Lemma 1, dictate that $\varepsilon$ chooses the interpretation of minimal cardinality, such that $[\beta_{\varphi, \mu}] = \varepsilon \odot (f_a(\mu) \lor f_b(\mu)) = \{\emptyset, \{c\}\}$. The result obtained, i.e., interpretation $c$, is the result of flipping the atom $a$ in the interpretation $ac \in [\mu]$; to recover $ac$ from $c$, we ‘reverse’ the original flips: one flip by $a$ and one by $b$, to get $[f_a(\beta_{\varphi, \mu}) \lor f_b(\beta_{\varphi, \mu})] = \{ac, bc\}$. By postulate $R_{\text{BGB}}$, we have that $[\varphi \odot \mu] = [f_a(\beta_{\varphi, \mu}) \lor f_b(\beta_{\varphi, \mu})] = \{ac\}$, i.e., exactly the result produced by the Dalal operator $\odot_{\text{H}, \text{min}}$.

## 5 Characterizing the Hamming Distance Min-Max Operator

The postulates put forward in Section 4 for characterizing the Dalal operator prove their worth in an additional sense, as they can be put to use, with minimal modifications, in characterizing the Hamming distance min-max operator $\odot_{\text{H}, \text{max}}$. This is the topic of the current section.

Of the newly proposed postulates, the Neutrality, Addition and Flipping postulates ($R_b$, $R_6$ and $R_T$, respectively) can be used as stated in Section 4, while the Best-Of-Best postulate $R_{\text{BGB}}$ has to be modified. Intuitively, this makes sense: postulates $R_b$, $R_6$ and $R_T$ are used in regulating what happens when the prior information is a complete formula $\varphi$, (alternatively, for what happens in the ranking that corresponds to the $\nu$-column in the table of distances, e.g., Table 1), in which case, as per Proposition 2, all operators presented here coincide, whereas postulate $R_{\text{BGB}}$ instructs us how to choose when the prior information consists of more than one model (alternatively, across different columns of the table of distances). Correspondingly, postulate $R_{\text{BGB}}$ encodes the constraint that revision should pick the best of the best models across all of the $\varphi_v$’s, for $v \in [\varphi]$, but this is not the rule that defines operator $\odot_{\text{H}, \text{max}}$. For $\odot_{\text{H}, \text{max}}$ we need a principle that mandates picking the best of the worst models across the $\varphi_v$’s. The key fact allowing us to do this relies on a certain duality specific to the Hamming distance.

![Figure 1: By flipping the atoms of $v \in [\varphi]$ in a model $w$ of $\mu$ we get an interpretation $w \triangle v$ whose size corresponds to the Hamming distance between $v$ and $w$.](image.png)
Proof. Thus, using Equality 1, \( d_\text{H}(v, w) = n - d_\text{H}(A \setminus v, w). \)

Intuitively, Lemma 4 implies that the further away \( w \) is from \( v \) (in terms of Hamming distance), the closer \( w \) is to \( A \setminus v \). In particular, we can infer that:

\[
d_\text{H}(v, w) = d_\text{H}(\emptyset, |v \triangle w|) = d_\text{H}(\emptyset, f_\emptyset(w)) = n - d_\text{H}(A, f_\emptyset(w)). \tag{1}
\]

Hence, \( w \in [\mu] \) is among the models of \( \mu \) at maximal Hamming distance to \( v \) if and only if \( f_\emptyset(w) \) is, among the models of \( f_\emptyset(\mu) \), the closest model to \( A \). Thus, we can define the best-of-worst formula \( \gamma_{\varphi, \mu} \) with respect to \( \varphi \) and \( \mu \) as:

\[
\gamma_{\varphi, \mu} = \varepsilon \circ \left( \bigvee_{v \in [\varphi]} f_\emptyset(\gamma_{\varphi, \mu}) \right) \cap \mu,
\]

i.e., as the result of revising the null formula \( \varepsilon \) by a disjunction made up of the results obtained from a sequence of revisions of the full formula \( \alpha \). In this sequence \( \alpha \) is revised, in turn, by \( f_\emptyset(\mu) \), for every model \( v \in [\varphi] \).

Thus, similarly as for \( \beta_{\varphi, \mu} \), from Section 4, \( \gamma_{\varphi, \mu} \) simulates the process of going through the table of Hamming distances (e.g., Table 1), except that in this case we are interested in (i) selecting the worst elements according to each \( \varphi \), for \( v \in [\varphi] \), an operation reflected by the revision \( \alpha \circ f_\emptyset(\mu) \), and (ii) selecting the best among these worst elements, an operation reflected by submitting the results obtained previously to \( \varepsilon \) for an additional round of revision. A bespoke postulate, called the Best-Of-Worst postulate \( R_{\text{BOW}} \), recovers the models of \( \mu \) from the models of \( \gamma_{\varphi, \mu} \):

\[
(R_{\text{BOW}}) \quad \varphi \circ \mu \equiv \left( \bigvee_{v \in [\varphi]} f_\emptyset(\gamma_{\varphi, \mu}) \right) \cap \mu.
\]

Postulate \( R_{\text{BOW}} \) stipulates that the result of revising \( \varphi \) by \( \mu \) consists of those models of \( \mu \) that come out of flipping \( \gamma_{\varphi, \mu} \), by each model of \( \varphi \), in this way reversing the initial flips that delivered the revision formula posed to \( \varepsilon \).

The proof that the postulates put forward actually characterize the \( \sigma_{\text{H}, \text{max}} \) operator hinges on \( \gamma_{\varphi, \mu} \) selecting interpretations corresponding to models of \( \mu \) that minimize maximal Hamming distance to \( \varphi \).

Lemma 5. If \( \circ \) is a revision operator that satisfies postulates \( R_1, R_3-R_6, R_8, R_4 \) and \( R_{\text{BOW}} \), then, for any formulas \( \varphi \) and \( \mu \) and interpretations \( w \) and \( v \), it holds that \( w \triangle v \in [\gamma_{\varphi, \mu}] \) if and only if \( w \in \arg\min_{w \in [\mu]} \max_{v \in [\varphi]} d_\text{H}(v, w) \).

Proof. Using postulates \( R_8, R_4 \) and \( R_6 \) we can prove that \( \alpha \) selects the models of \( \mu \) that minimize Hamming distance to \( A \), in a way completely analogous to Lemmas 1 and Lemma 2. Thus, using Equality 1, \( \alpha \circ f_\emptyset(\mu) \) selects interpretations \( w \triangle v \) such that \( d_\text{H}(v, w) = \max_{w \in [\mu]} d_\text{H}(v, w) \). Then, using Lemma 1, we obtain that \( \gamma_{\varphi, \mu} \) selects interpretations \( w \triangle v \) where \( w \) minimizes max-distance to \( \varphi \).

With Lemma 5 the characterization of \( \sigma_{\text{H}, \text{max}} \) follows immediately.

Theorem 2. If \( \circ \) is a revision operator, then \( \circ \) satisfies postulates \( R_1, R_3-R_6, R_8, R_4 \) and \( R_{\text{BOW}} \) iff \( \circ \equiv \sigma_{\text{H}, \text{max}} \).

The proof is similar, in its essentials, to the proof of Theorem 1 and is therefore omitted. The following example, however, illustrates how the mechanism works on a concrete case.

Example 5. Consider formulas \( [\varphi] = \{a, b\} \) and \( [\mu] = \{a, b, c\} \), as in Example 4, over the set \( A = \{a, b, c\} \) of atoms. Using the \( \sigma_{\text{H}, \text{max}} \) operator we obtain that \( [\varphi] \circ_{\text{H}, \text{max}} [\mu] = \{ab\} \), but we can show that a (putatively different) revision operator \( \circ \) known only to satisfy the stated postulates arrives at the same conclusion. It does so by first figuring out, using postulates \( R_1, R_3-R_6, R_8, R_4 \) that \( [\alpha \circ f_\emptyset(\mu)] = \{bc\} \) and \( [\alpha \circ f_\emptyset(\mu)] = \{ab\} \), with \( \alpha \), in this case, such that \( [\alpha] = \{ab\} \) (see Figure 2 for an illustration). At this point, we have obtained the (flipped versions of) the models of \( \mu \) at maximal Hamming distance to \( a \) and \( b \), respectively. Following this, we get that \( [\gamma_{\varphi, \mu}] = [\varepsilon \circ ((\alpha \circ f_\emptyset(\mu) \vee (\alpha \circ f_\emptyset(\mu))) = \{bc\} \), where \( bc \) was obtained from \( abc \) by flipping \( a \). Postulate \( BOW \) then be recovers \( abc \) through an extra flip of \( a \).

6 Characterizing the Hamming Surprise Min-Max Operator

Finally, we return to the operator \( \sigma_{\text{H}, \text{max}} \) and, using the wisdom gained in Section 4 and 5, provide it with an axiomatic foundation. In doing so we pursue that same strategy as in the previous sections: (i) establish, axiomatically, what the revision result should be in the ‘base’ case in which the prior belief is of a simple type, which can be decided by appeal to an argument using appealing notions of symmetry; (ii) reduce, axiomatically, an arbitrary instance \( \varphi \circ \mu \) of revision to the base case, in a manner that preserves the result of \( \sigma_{\text{H}, \text{max}} \) on the given instance.

The base case for this section consists, as for the \( \sigma_{\text{H}, \text{max}} \) operator, of revision when prior information is either \( \varepsilon \) or \( \alpha \), and we want to make sure we employ a set of postulates...
that deliver the expected result: since \( o^{\text{SH}, \max} \) behaves exactly like the Dalal and \( o^{\text{SH}, \max} \) operators when prior information is complete, postulates \( R_8, R_9 \) and \( R_6 \) can be used without modification (the assumption of completeness made in Section 4 pays off here). We can also use the standard postulates \( R_1 \) and \( R_3-R_4 \), which we already know \( o^{\text{SH}, \max} \) satisfies (see Proposition 1). Postulates \( R_5-R_6 \) are, however, problematic, since \( o^{\text{SH}, \max} \) does not satisfy them in their unrestricted form (also Proposition 1). However, the equivalence of \( o^{\text{SH}, \max} \) with the Dalal and \( o^{\text{SH}, \max} \) operators when prior information is complete means that we can use postulates \( R_5 \) and \( R_6 \), restricted to the case when \( \varphi \) is complete. The restrictions are denoted \( R_5^c \) and \( R_6^c \), respectively.

The next step involves engineering a choice situation focused on \( \alpha \) and \( \varepsilon \) that is equivalent, in terms of what gets chosen, to the mechanics of \( o^{\text{SH}, \max} \). This is done using a few intermediary notions, as follows. If \( \varphi \) and \( \mu \) are formulas such that \( [\varphi] = \{v_1, \ldots, v_n\} \), the adjunction interpretations \( x_1, \ldots, x_n \) are interpretations consisting of completely new atoms such that \( |x_i| = d_H(v_i, \mu) \). For \( v_i \in [\varphi] \), the corrected interpretation \( v_i^* \) is defined as \( v_i^* = v_i \cup \{x_1 \cup x_{i-1} \cup x_{i+1} \cup \cdots \cup x_n\} \), i.e., as the result of adding to \( v_i \) all the adjunction interpretations, except \( x_i \). Then, the best-surprise formula \( \sigma_{\varphi, \mu} \) with respect to \( \varphi \) and \( \mu \) is defined as:

\[
\sigma_{\varphi, \mu} = \varepsilon \circ \left( \bigvee_{v_i \in [\varphi]} (\alpha \circ f_{v_i^*}(\mu)) \right).
\]

In words, inside the main parenthesis we repeatedly revise \( \alpha \) by a flipped version of \( \mu \); one revision for every model \( v_i \) of \( \varphi \), flipping \( \mu \) by the atoms in the corrected interpretation \( v_i^* \). The disjunction of all these revisions is then passed on to \( \alpha \) for another round of revision.

The reasoning behind this definition is that it recasts the surprise min-max revision scenario for \( [\varphi] = \{v_1, \ldots, v_n\} \) and \( \mu \) into a min-max distance revision scenario for \( [\varphi^*] = \{v_1^*, \ldots, v_n^*\} \) and \( \mu \) (which we know how to axiomatize from Section 5), while keeping the relative ranking of the models of \( \mu \) intact. The result following makes this precise.

**Lemma 6.** If \( \varphi \) and \( \mu \) are propositional formulas, \( v_i, v_k \in [\varphi] \) and \( w_j, w_k \in [\mu] \), then \( s_H^\mu(v_i, w_j) \leq s_H^\mu(v_k, w_k) \) iff \( d_H(v_i^*, w_j) \leq d_H(v_k^*, w_k) \).

**Proof.** Take \( [\varphi] = \{v_1, \ldots, v_n\} \), and \( m_i = d_H(v_i, \mu) \), for \( v_i \in [\mu] \). We have that:

\[
s_H^\mu(v_i, w_j) \leq s_H^\mu(v_k, w_k) \iff d_H(v_i^*, w_j) \leq d_H(v_k^*, w_k) = m_k.
\]

We now add \( \sum_{1 \leq r \leq n} m_r \) on both sides, to get an equivalence with \( d_H(v_i, w_j) + \sum_{1 \leq r \leq n, r \neq k} m_r \leq d_H(v_k, w_k) + \sum_{1 \leq r \leq n, r \neq k} m_r \). This, in turn, is equivalent to \( d_H(v_i \cup (\bigcup_{1 \leq r \leq n, r \neq k} x_r), w_j) \leq d_H(v_k \cup (\bigcup_{1 \leq r \leq n, r \neq k} x_r), w_k) \), which can be rewritten as \( d_H(v_i^*, w_j) \leq d_H(v_k^*, w_k) \). \( \square \)

Intuitively, the table of Hamming distances for \( [\varphi^*] = \{v_1^*, \ldots, v_n^*\} \) and \( \mu \) can be thought of as obtained from the surprise table for \( \varphi \) and \( \mu \) (see, e.g., Table 2) by adding a constant term (i.e., \( \sum_{1 \leq r \leq n} m_r \)) to every entry, a transformation that does not modify the relationships between the values: the \( v_i^* \) are the interpretations that induce the appropriate distances. This ensures that the models of \( \sigma_{\varphi, \mu} \) obtained through a min-max distance type of postulate, correspond to models of \( \mu \) that minimize maximum surprise with respect to \( \varphi \) and relative to \( \mu \), and warrants the following postulate, called **Best-of-Worst-Surprise:**

\[
(R_{\text{BOWS}}) \quad \varphi \circ \mu \equiv \left( \bigvee_{v \in [\varphi]} f_{v^*}(\sigma_{\varphi, \mu}) \right) \land \mu.
\]

As expected, the \( R_{\text{BOWS}} \) postulate delivers exactly those models of \( \mu \) that minimize maximum surprise, and underpins the final characterization result.

**Theorem 3.** A revision operator \( \circ \) satisfies postulates \( R_1, R_3-R_4, R_5^c-R_6^c, R_8, R_9, R_{\text{BOWS}} \) iff \( \equiv \equiv o^{\text{SH}, \max} \).

The following example illustrates the way in which postulate \( R_{\text{BOWS}} \) obtains the revision result.

**Example 6.** Consider, again, formulas \( [\varphi] = \{a, b\} \) and \( [\mu] = \{ac, abc\} \). We have that \( \{v^*, \varphi \} \equiv \{ac, abc\} \). Assuming we are working with an operator \( \circ \) of which the only thing we know is that it satisfies the postulates in Theorem 3, we notice that \( d_H(a, \mu) = 1 \) and \( d_H(b, \mu) = 2 \). The postulates then direct us to compute the Hamming distance min-max result for \( [\varphi^*] = \{ac, bc\} \) and \( \mu \), with \( x \) and \( yz \) as the adjunction interpretations. The result obtained in this way is exactly \( \{ac, abc\} \).

**7 Conclusion**

We have introduced the Hamming surprise min-max operator \( o^{\text{SH}, \max} \), a revision operator that minimizes surprise relative to the prior belief as well as the newly acquired information. We have shown that, even though \( o^{\text{SH}, \max} \) does not satisfy all standard KM revision postulates, it is underpinned, in its choice behavior, by principles similar to those guiding established revision operators, among them appealing symmetry notions such as invariance under renamings and flips. When unearthed and formulated as logical postulates, these principles (or slight variations thereof) turned out to be powerful enough to fully characterize not just the surprise operator, but also the existing Dalal and Hamming distance min-max operator.

One obvious direction for future work lies in taking the idea of context dependence further: what other aspects of the environment influence an agent’s plausibility rankings? Things that come to mind are issues of trust, the ‘strangeness’ of the new information, or peer effects. An alternative is to exploit the bottom-up, DIY nature of some of the postulates presented here in order to construct a framework, similar to that employed in collective decision-making (Cailloux and Endriss 2016), for offering justifications for revision results, i.e., human-readable and at the same time rigorous step-by-step arguments for how to obtain a particular result, starting from a specific set of postulates. Finally, the assumptions embedded in the present treatment call for taking the epistemic stance seriously, and investigating the relative worth of the various revision operators with respect to recovering the ground truth.
References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. J. Symb. Log. 50(2):510–530.

Aravanis, T. I.; Peppas, P.; and Williams, M. 2021. An investigation of parametrized difference revision operators. Ann. Math. Artif. Intell. 89(1-2):7–28.

Cailloux, O., and Endriss, U. 2016. Arguing about Voting Rules. In Jonker, C. M.; Marsella, S.; Thangarajah, J.; and Tuyls, K., eds., Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016, 287–295. ACM.

Dalal, M. 1988. Investigations into a Theory of Knowledge Base Revision. In Proceedings of the 7th National Conference on Artificial Intelligence, 1988, 475–479.

Fermé, E. L., and Hansson, S. O. 2018. Belief Change: Introduction and Overview. Springer Briefs in Intelligent Systems. Springer.

Hansson, S. O. 2017. Logic of Belief Revision. In Zalta, E. N., ed., The Stanford Encyclopedia of Philosophy. Winter 2017 edition.

Herzig, A., and Rifi, O. 1999. Propositional Belief Base Update and Minimal Change. Artificial Intelligence 115(1):107–138.

Hohwy, J. 2016. The Self-Evidencing Brain. Noûs 50(2):259–285.

Katsuno, H., and Mendelzon, A. O. 1992. Propositional Knowledge Base Revision and Minimal change. Artificial Intelligence 52(3):263–294.

Lave, C. A., and March, J. G. 1993. An Introduction to Models in the Social Sciences. University Press of America.

Marquis, P., and Schwind, N. 2014. Lost in translation: Language independence in propositional logic-application to belief change. Artificial Intelligence 206:1–24.

Milnor, J. 1954. Games against nature. In Thrall, R.; Coombs, C.; and Davis, R., eds., Decision Processes. New York: Wiley.

Parikh, R. 1999. Beliefs, Belief Revision, and Splitting Languages. Logic, Language and Computation 2(96):266–268.

Peppas, P., and Williams, M. 2016. Kinetic Consistency and Relevance in belief revision. In Michael, L., and Kakas, A. C., eds., Proceedings of JELIA 2016, volume 10021 of Lecture Notes in Computer Science, 401–414.

Peppas, P., and Williams, M. 2018. Parametrised Difference Revision. In Thielscher, M.; Toni, F.; and Wolter, F., eds., Proceedings of KR 2018, 277–286. AAAI Press.

Peppas, P.; Williams, M.; Chopra, S.; and Foo, N. Y. 2015. Relevance in belief revision. Artif. Intell. 229:126–138.

Peppas, P. 2008. Belief Revision. In van Harmelen, F.; Lifschitz, V.; and Porter, B. W., eds., Handbook of Knowledge Representation, volume 3. Elsevier. 317–359.

Peterson, M. 2017. An Introduction to Decision Theory. Cambridge University Press, Second edition.

Pozos-Parra, P.; Liu, W.; and Perrussel, L. 2013. Dalal’s Revision without Hamming Distance. In Mexican International Conference on Artificial Intelligence, 41–53. Springer.

Rott, H. 1992. Modellings for Belief Change: Base Contrac
tion, Multiple Contraction, and Epistemic Entrenchment. In Proceedings of JELIA ’92, 139–153.

Sen, A. 1993. Internal Consistency of Choice. Econometrica 61(3):495–521.

Sen, A. K. 2017. Collective Choice and Social Welfare: Expanded Edition. Penguin UK.