Revealed Epistemic Trust

Xu Li[1] , **Leendert van der Torre**[1] , **Liuwen Yu**[2]
[1]University of Luxembourg
[2]Luxembourg Institute of Science and Technology
{xu.li, leon.vandertorre}@uni.lu, liuwen.yu@list.lu

Abstract

Inspired by revealed preference in economics, we study *revealed epistemic trust*: an agent's (dis)trust in an information source is typically hidden, while her accept/reject behavior leaves observable traces. We model such traces by an *acceptance function* that maps each reported set of formulas to the subset the agent accepts. We develop two complementary models: a *white-list* mode, where acceptance is supported by trusted information in the report, and a *black-list* mode, where acceptance avoids distrusted patterns via a cautious remainder-set/full-meet construction. For both modes, we provide postulate-based representation theorems and show how canonical "revealed" trust and distrust cores can be reconstructed from the acceptance function itself.

1 Introduction

Revealed preference theory in economics starts from a methodological asymmetry: preferences are private, but choices are observable. It asks when observed choice behavior can be *rationalized* by some underlying preference structure, and it characterizes rationalizable behavior via axioms on the choice function.

We adopt the same stance for *epistemic trust*. In many information settings, an agent's (dis)trust in a source is hidden. What we can often observe, however, are public traces of how the agent handles communicated information—for instance, in a dialogue she may accept a claim, reject it, or suspend judgment. We treat such traces as *revealed belief* (or revealed commitment) and abstract them by an *acceptance function* $\alpha : \wp(\mathcal{L}) \to \wp(\mathcal{L})$, where $\alpha(A)$ is the subset of a reported set A that the agent accepts (typically $\alpha(A) \subseteq A$). Of course, public traces typically provide only a very partial view of what the agent might accept, unlike an "acceptance function" which provides a complete description.[1] An acceptance function is an idealized profile, analogous to a choice function in revealed preference theory. Our results can be used to assess whether the observable public traces are consistent with a given acceptance mode.

We consider two acceptance modes: *white-list* (accept only what is trusted) and *black-list* (accept everything except what is ruled out by distrust). For example, in a dialogue a consultant's bare recommendation is not accepted until supporting information is provided, and the agent's replies thereby constitute revealed belief (Example 1 in Section 2). This is white-list behavior: a claim is accepted only when it comes with enough trusted support. By contrast, a customer may accept each sales claim alone but withhold both when the claims are combined, again revealing belief through a black-list "red-flag" reaction (Example 2 in Section 2).

In this paper, we answer the following research questions.

1. When can an observed acceptance function α be explained by a hidden set T of trusted statements, so that α behaves like "accept what is supported by trusted information in the report"?
2. When can α be explained by a hidden set D of distrusted statements, so that α behaves like "accept everything except what cannot be kept without triggering distrust" (and why does naive deletion fail)?
3. If such a T or D exists, can we reconstruct a natural "revealed" rationalizer from α alone?
4. Which simple postulates on α characterize each mode, and which postulates separate white-lists from black-lists?
5. How do standard closure constraints on (dis)trust (e.g., equivalence-, conjunction-, disjunction-, negation-, or consequence-closure) change what acceptance behavior is representable?

We proceed in a revealed-structure style. First, we treat α as the observable object and introduce two generative models: *rationalizability* and *b-rationalizability*. Second, for each model, we prove a representation theorem: a small set of postulates on α is necessary and sufficient for membership in the corresponding class. Third, we reconstruct canonical "revealed" rationalizers from behavior: $Fix(\alpha)$ yields a maximal trusted rationalizer in the white-list case, and $Dis(\alpha)$ yields a maximal distrusted rationalizer in the black-list case. Finally, we study constrained variants obtained by imposing closure conditions on T and D.

The layout of this paper is as follows. Section 2 recalls propositional preliminaries and acceptance functions. Section 3 develops the white-list approach (rationalizability, representation, and $Fix(\alpha)$, plus constrained variants). Section 4 develops the black-list approach (b-rationalizability

[1]We thank an anonymous reviewer for pointing this out.

via remainder sets, representation, and $Dis(\alpha)$, plus constrained variants). Section 5 discusses connections to logics of trust, abstract choice theory, and belief revision. Section 6 concludes and outlines open problems.

2 Preliminaries

In this section, we recall definitions and facts from classical propositional logic, which will be used in later sections.

Let PROP be a nonempty (finite or countably infinite) set of *propositional variables* or *atoms*. $\mathcal{L}$ is the propositional language generated by PROP, and the elements of $\mathcal{L}$ are called *formulas*. We define $\top$ as an abbreviation of $p \vee \neg p$ (for a certain atom p) and $\bot := \neg\top$. Given a finite set A of formulas, $\bigwedge A$ denotes the conjunction of all formulas in A (note that $\bigwedge \emptyset := \top$).

For all sets of formulas $A \cup \{a\}$, $A \vdash a$ denotes that a is a logical consequence of A in classical propositional logic. $Cn(A) = \{a \mid A \vdash a\}$ is the set of all the logical consequences of A. We will omit the braces when we explicitly list the elements of A. That is, instead of $\{a\} \vdash b$ and $Cn(\{a, b\})$, we write $a \vdash b$ and $Cn(a, b)$. In addition, for two formulas a and b, $a \dashv\vdash b$ abbreviates the conjunction of $a \vdash b$ and $b \vdash a$. The well-known properties of the operator Cn are taken for granted, such as Inclusion, Monotony, Idempotence, and Compactness.

Definition 1. A function $\alpha : \wp(\mathcal{L}) \to \wp(\mathcal{L})$ is called an *acceptance function*.

Intuitively, given reports A from a source, $\alpha(A)$ specifies which statements in A will be accepted by an agent. The agent may not accept all the statements in A, because she may not trust all of them, or she may distrust some of them. Thus, we have $\alpha(A) \subseteq A$ in general, but the equality may fail. As before, the braces of A will be omitted if its elements are explicitly given. I.e., we write $\alpha(a, b)$ instead of $\alpha(\{a, b\})$.

In this paper, we consider two acceptance modes of the agent: the white-list and black-list. In the white-list approach, the agent trusts the source on a set of statements and accepts only what is trusted. On the contrary, in the black-list approach, the agent distrusts the source on a set of statements and accepts anything except those that are distrusted.

Two motivating examples are provided below (both adapted from (Liau 2003)). They illustrate two common patterns of selective acceptance: in the first, a recommendation is accepted only when accompanied by supporting information; in the second, each claim is acceptable on its own, but their conjunction triggers a "red-flag" reaction and leads to withholding both.

Example 1. Consider a conversation between a financial consultant and a decision agent. The consultant first tells the agent that "it is worthwhile to invest in company X." The agent does not accept this. The consultant then adds that "the financial situation of company X is excellent." The agent now accepts both statements. Let p mean "it is worthwhile to invest in company X" and let q mean "the financial situation of company X is excellent." This behavior can be represented by an acceptance function α such that $\alpha(p) = \emptyset$ and $\alpha(p, q) = \{p, q\}$.

Example 2. Consider a conversation between a salesman and a customer in a shop. The salesman tells the customer that "the camera has a low price." The customer accepts this. The salesman then adds that "it is of high quality." The customer becomes skeptical, and withholds judgment (e.g., replying "Thanks, I'll think about it"). Let p and q denote "the camera has a low price" and "the camera is of high quality," respectively. This behavior can be represented by an acceptance function β such that $\beta(p) = \{p\}$ and $\beta(p, q) = \emptyset$.

As we shall show later, the acceptance function α is only "rationalizable" in the white-list sense, whereas β is only "rationalizable" in the black-list sense.

Remark 1. An acceptance function α is a choice function on the domain of propositional language, see Section 5. As in the abstract choice theory literature (Sen 1971; Chambers and Echenique 2016), we may consider restrictions on the domain of α. For example, α may only accept finite sets as inputs because the reports from the source are typically finite in reality. On the other hand, the inputs of α may be logically closed sets of formulas, as in the AGM theory of belief revision (Alchourrón, Gärdenfors, and Makinson 1985). These restrictions on the domain have non-trivial impacts on our representation results to be introduced. However, we choose to start from the most general form of acceptance functions and leave these interesting directions for future research.

3 The White-list Approach

In this section, we study the white-list acceptance mode, i.e., the agent accepts only what is trusted. We first define the notion of "rationalizability" of an acceptance function, which is intended to formalize the white-list acceptance mode of the agent. Then we focus on the task of finding a set of properties of acceptance functions that completely characterize the class of all rationalizable acceptance functions. Our main result is a representation theorem. Finally, we study how the properties of epistemic trust can affect the properties of acceptance functions.

Given a set T of formulas consisting of all statements trusted by the agent (i.e., $a \in T$ means that "the agent trusts the source on a"), and a set A of informed statements, which parts of A will be accepted by the agent? The definition of "rationalizability" below states that the agent accepts only the elements of A that can be inferred from the trusted information in A.

Definition 2 (Rationalizability)**.** A set T of formulas *rationalizes* an acceptance function α if $\alpha(A) = Cn(T \cap Cn(A)) \cap A$ for all sets A of formulas. An acceptance function α is *rationalizable* if there exists a set T of formulas rationalizing α.

In the above definition, T is not required to be closed under logical consequences, because Liau (2003) has already argued that epistemic trust is not closed under logical consequences. At this stage, we assume no properties of T. We will study more properties of T later.

Definition 2 is illustrated by the following example:

Example 3 (Example 1 continued)**.** Let $T = \{p \wedge q\}$ and α be rationalized by T. Then $p \in \alpha(p,q) = Cn(T \cap Cn(p,q)) \cap \{p,q\} = \{p,q\}$. However, $p \notin \alpha(p) = Cn(T \cap Cn(p)) \cap \{p\} = \emptyset$.

Our next goal is to find a set of properties that completely characterize the class of all rationalizable acceptance functions. Below we list some properties of acceptance functions:

Exclusion	$\alpha(A) \subseteq A$.
Right weakening	$Cn(\alpha(A)) \cap A \subseteq \alpha(A)$.
Left strengthening	$\alpha(Cn(A)) \cap A \subseteq \alpha(A)$.
Monotony	If $B \subseteq A$ then $\alpha(B) \subseteq \alpha(A)$.
Groundedness	If $x \in \alpha(A)$ then there is a finite $B \subseteq Cn(A)$ s.t. $B \cup \{x\} \subseteq \alpha(B)$.

Exclusion is the inverse of the well-known Inclusion property. Right weakening states that if an element of A can be inferred from what are accepted in A, then it must also be accepted (i.e., the accepted formulas are closed under logical consequences). Left strengthening says that if an element of A is accepted when all the consequences of A are given, then it must be accepted given only A.

From the first three properties, we can derive the following Reasoning, which will simplify some later proofs.

Reasoning $\quad Cn(\alpha(Cn(A))) \cap A \subseteq \alpha(A)$.

Propostion 3. *For all acceptance functions α, if α satisfies* Exclusion*,* Right weakening*, and* Left strengthening*, then it satisfies* Reasoning.

Proof. Suppose α satisfies the mentioned three properties. By Right weakening, $Cn(\alpha(Cn(A))) \cap Cn(A) \subseteq \alpha(Cn(A))$. Since $\alpha(Cn(A)) \subseteq Cn(A)$ (by Exclusion), $Cn(\alpha(Cn(A))) \subseteq Cn(Cn(A)) = Cn(A)$. Thus, $Cn(\alpha(Cn(A))) = Cn(\alpha(Cn(A))) \cap Cn(A) \subseteq \alpha(Cn(A))$. Since $\alpha(Cn(A)) \cap A \subseteq \alpha(A)$ by Left strengthening, $Cn(\alpha(Cn(A))) \cap A \subseteq \alpha(A)$. □

The Monotony property is self-evident. But Groundedness needs more explanation. Intuitively, it says that if x is accepted given A, then there must be a ground B for the acceptance of x. The ground may not be x itself, but it may contain more information. For instance, in Example 3 we have $p \in \alpha(p,q)$. The ground for the acceptance of p may be $\{p,q\}$ because $\{p,q\} \subseteq \alpha(p,q)$, but p itself cannot be a ground because $p \notin \alpha(p)$.

Another way to understand Groundedness is to note that, when PROP is finite and all other properties are present, Groundedness is equivalent to the following Restricted Idempotence:

R-Idempotence $\quad \alpha(A) \subseteq \alpha(\alpha(A))$, if $A = Cn(A)$.

Propostion 4. *The following hold for any α:*

(1) If α satisfies Monotony *and* Groundedness*, then α satisfies* R-Idempotence*.*

(2) Let PROP *be finite and α satisfy* Exclusion*,* Right weakening*,* Left strengthening*, and* Monotony*. Then α satisfies* Groundedness *if and only if it satisfies* R-Idempotence*.*

Proof. (1) Suppose that α satisfies Monotony and Groundedness, and $A = Cn(A)$. If $x \in \alpha(A)$, then there is $B \subseteq Cn(A)$ such that $B \cup \{x\} \subseteq \alpha(B)$. Since $B \subseteq Cn(A) = A$, $\alpha(B) \subseteq \alpha(A)$ by Monotony. Thus, $B \subseteq \alpha(B) \subseteq \alpha(A)$. Thus, by Monotony, $\alpha(B) \subseteq \alpha(\alpha(A))$. Therefore, $x \in \alpha(B) \subseteq \alpha(\alpha(A))$.

(2) Suppose PROP is finite and α satisfies the mentioned properties. Then, by Proposition 3, α also satisfies Reasoning. The direction from left to right follows from (1).

From right to left. Suppose that α satisfies R-Idempotence. Let $x \in \alpha(A)$. Note that, since PROP is finite, there must be a finite set B' of formulas such that $Cn(B') = Cn(\alpha(Cn(A)))$. Let $B = B' \cup \{x\}$. Since $x \in \alpha(Cn(A))$ (by Monotony), we have $Cn(B) = Cn(\alpha(Cn(A)))$. To show that α satisfies Groundedness, it suffices to show the following properties of B:

$$(1)\ B \subseteq Cn(A), \text{ and } (2)\ B \subseteq \alpha(B).$$

For (1), note that $B \subseteq Cn(\alpha(Cn(A)))$. Since $\alpha(Cn(A)) \subseteq Cn(A)$ by Exclusion, $Cn(\alpha(Cn(A))) \subseteq Cn(Cn(A)) = Cn(A)$. Therefore, $B \subseteq Cn(A)$.

For (2), let $y \in B$. Then $y \in Cn(\alpha(Cn(A)))$. By R-Idempotence, we have $\alpha(Cn(A)) \subseteq \alpha(\alpha(Cn(A)))$. Since $\alpha(Cn(A)) \subseteq Cn(B)$, $\alpha(\alpha(Cn(A))) \subseteq \alpha(Cn(B))$ by Monotony. Therefore, $\alpha(Cn(A)) \subseteq \alpha(Cn(B))$ and, thus, $Cn(\alpha(Cn(A))) \subseteq Cn(\alpha(Cn(B)))$. Therefore, $y \in Cn(\alpha(Cn(B)))$. Since $y \in B$, by Reasoning it follows that $y \in \alpha(B)$. □

Remark 2. If PROP is infinite, then R-Idempotence does not imply Groundedness, even when we assume all other properties. We call a set A of formulas "finitely based" if there is a finite B such that $A \subseteq Cn(B)$. Consider the following acceptance function α:

$$\alpha(A) = \begin{cases} Cn(\emptyset) \cap A, & \text{if } A \text{ is finitely based;} \\ A, & \text{otherwise.} \end{cases}$$

It can be verified that α satisfies Exclusion – Monotony. Moreover, it satisfies R-Idempotence: We show that $\alpha(Cn(A)) \subseteq \alpha(\alpha(Cn(A)))$. If $Cn(A)$ is finitely based, then $\alpha(Cn(A)) = Cn(\emptyset) \cap Cn(A) = Cn(\emptyset) = \alpha(Cn(\emptyset)) = \alpha(\alpha(Cn(A)))$. If $Cn(A)$ is not finitely based, then $\alpha(Cn(A)) = Cn(A) = \alpha(\alpha(Cn(A)))$.

However, α does not satisfy Groundedness: We have $p \in \alpha(\text{PROP})$. But for all finite $B \subseteq Cn(A)$, since B is finitely based, $\alpha(B) = Cn(\emptyset) \cap B$. Therefore, $p \notin \alpha(B)$.

R-Idempotence is a restricted form of the following Idempotence. However, it can be shown that a rationalizable acceptance function may not satisfy Idempotence.

Idempotence $\quad \alpha(A) \subseteq \alpha(\alpha(A))$.

Example 4. There exists a rationalizable acceptance function α which does not satisfy Idempotence. Let $\alpha(A) = Cn(\{a \wedge b\} \cap Cn(A)) \cap A$. The following holds:

- $\alpha(a \wedge b \wedge c, a) = \{a\}$.
- $\alpha(a) = \emptyset$.

Next, we verify that Exclusion – Groundedness are sound for rationalizable acceptance functions.

Propostion 5 (Soundness)**.** *Let α be any rationalizable acceptance function. Then α satisfies* Exclusion*,* Right weakening*,* Left strengthening*,* Monotony*, and* Groundedness*.*

Proof. Suppose that α is rationalized by T. The cases for Exclusion and Monotony are trivial.

Right weakening. Note that $Cn(T \cap Cn(A)) \cap A \subseteq Cn(T \cap Cn(A))$. Thus, $Cn(\alpha(A)) = Cn(Cn(T \cap Cn(A)) \cap A) \subseteq Cn(T \cap Cn(A))$. Therefore, $Cn(\alpha(A)) \cap A \subseteq Cn(T \cap Cn(A)) \cap A = \alpha(A)$.

Left strengthening. Note that $\alpha(Cn(A)) \cap A = Cn(T \cap Cn(A)) \cap Cn(A) \cap A = \alpha(A)$.

Groundedness. Suppose $x \in \alpha(A) = Cn(T \cap Cn(A)) \cap A$. Then, by the compactness of Cn, there must be $a_1, \ldots, a_n \in T \cap Cn(A)$ such that $a_1, \ldots, a_n \vdash x$. Let $B = \{a_1, \ldots, a_n, x\}$. It is easy to see that $B \cup \{x\} \subseteq \alpha(B)$. □

Remark 3. From the soundness result, it follows that the acceptance function β in Example 2 is not rationalizable, because Monotony fails for β.

It remains to show that Exclusion – Groundedness are sufficient for the rationalizability of acceptance functions. This is the so-called representation theorem. To do so, the next proposition is helpful. Given an acceptance function α, let $Fix(\alpha) = \{a \in \mathcal{L} \mid a \in \alpha(a)\}$.

Propostion 6. *An acceptance function α is rationalizable if and only if $Fix(\alpha)$ rationalizes α.*

Proof. We show only the "only if" part. Suppose α is rationalizable. Then $\alpha(A) = Cn(T \cap Cn(A)) \cap A$ for some set T of formulas. We need to show that $\alpha(A) = Cn(Fix(\alpha) \cap Cn(A)) \cap A$, i.e.,

$$Cn(T \cap Cn(A)) \cap A = Cn(Fix(\alpha) \cap Cn(A)) \cap A.$$

For the inclusion $\subseteq$, it suffices to show that $T \subseteq Fix(\alpha)$. This holds because for any formulas $a \in T$, $a \in \alpha(a)$. For the inclusion $\supseteq$, it suffices to show that $Fix(\alpha) \cap Cn(A) \subseteq Cn(T \cap Cn(A))$. Suppose $a \in Fix(\alpha) \cap Cn(A)$. Since $a \in \alpha(a)$, there are $x_1, \ldots, x_n \in T \cap Cn(a)$ such that $x_1, \ldots, x_n \vdash a$. Note that $x_1, \ldots, x_n \in T \cap Cn(A)$, since $a \in Cn(A)$. Hence, $a \in Cn(T \cap Cn(A))$. □

Note that $Fix(\alpha)$ is the inclusion-maximal set that rationalizes α, given that α is rationalizable. In addition, $Fix(\alpha)$ contains all tautologies and is closed under logical equivalences and conjunction.

Propostion 7. *Let α be a rationalizable acceptance function. Then the following holds:*

(1) $\top \in Fix(\alpha)$;

(2) If $a \in Fix(\alpha)$ and $a \dashv\vdash b$, then $b \in Fix(\alpha)$;

(3) If $a, b \in Fix(\alpha)$, then $a \wedge b \in Fix(\alpha)$.

Proof. We show only (3). Suppose α is rationalized by T and $a, b \in Fix(\alpha)$. Then $a \in Cn(T \cap Cn(a))$ and $b \in Cn(T \cap Cn(b))$. Thus, $a \wedge b \in Cn(T \cap (Cn(a) \cup Cn(b))) \subseteq Cn(T \cap Cn(a \wedge b))$. Thus, $a \wedge b \in \alpha(a \wedge b)$. □

Now we are ready to present the representation theorem.

Lemma 8. *Let α satisfy* Exclusion*,* Left strengthening*,* Right weakening*,* Monotony*, and* Groundedness*. Then $\alpha(A) = Cn(Fix(\alpha) \cap Cn(A)) \cap A$ for all A.*

Proof. Suppose that α satisfies the mentioned properties. Then, by Proposition 3, α also satisfies Reasoning. We show that $\alpha(A) = Cn(Fix(\alpha) \cap Cn(A)) \cap A$. We first show the inclusion $\subseteq$. By Exclusion it suffices to show that $\alpha(A) \subseteq Cn(Fix(\alpha) \cap Cn(A))$. Suppose $x \in \alpha(A)$. By Groundedness there is finite $B \subseteq Cn(A)$ such that $B \cup \{x\} \subseteq \alpha(B)$. Since $B \subseteq Cn(\bigwedge B)$, $\alpha(B) \subseteq \alpha(Cn(\bigwedge B))$ by Monotony. Therefore, $B \subseteq \alpha(Cn(\bigwedge B))$. Since $\bigwedge B \in Cn(\alpha(Cn(\bigwedge B)))$, by Reasoning it follows that $\bigwedge B \in \alpha(\bigwedge B)$. Thus, $\bigwedge B \in Fix(\alpha)$. Since $\bigwedge B \in Cn(A)$ and $x \in \alpha(B) \subseteq B$ (by Exclusion), it follows that $x \in Cn(Fix(\alpha) \cap Cn(A))$.

It remains to show that $\alpha(A) \supseteq Cn(Fix(\alpha) \cap Cn(A)) \cap A$. By the Monotony of Cn and Reasoning, it suffices to show that $Fix(\alpha) \cap Cn(A) \subseteq \alpha(Cn(A))$. Let $x \in Fix(\alpha) \cap Cn(A)$. Then $x \in \alpha(x)$. Since $x \in Cn(A)$, $\alpha(x) \subseteq \alpha(Cn(A))$ by Monotony. Hence, $x \in \alpha(Cn(A))$. □

Theorem 1 (Representation)**.** *If α satisfies* Exclusion*,* Left strengthening*,* Right weakening*,* Monotony*, and* Groundedness*, then α is rationalizable.*

In case we consider only finitely many atoms, Groundedness can be replaced by R-Idempotence and the same representation result holds.

Corollary 9. *Suppose* PROP *is finite. The following two statements are equivalent for any acceptance function α:*

(1) α is rationalizable;

(2) α satisfies Exclusion*,* Left strengthening*,* Right weakening*,* Monotony*, and* R-Idempotence*.*

Proof. The implication from (1) to (2) follows from Propositions 5 and 4. The converse follows from Theorem 1 and Proposition 4. □

3.1 Constraints on T

In the above, the trust set T used to rationalize an acceptance function is not required to have any properties. However, reasonable logical structures can be considered on T, e.g., if the agent trusts both p and q, then she will also trust their conjunction $p \wedge q$. In this subsection, we consider several closure properties on T, which are proposed as reasonable inference rules for epistemic trust in the literature. We study their impacts on the acceptance functions. As we shall see, some inference rules for trust have no impact on the properties of acceptance functions in certain cases.

The constraints on T are listed below. EQ means that epistemic trust is closed under logical equivalence. EQ is considered the only basic rule for trust inference in (Liau 2003). ST abbreviates "Symmetric Trust", which states that epistemic trust is closed under negation and is accepted by Liau (2003) as an optional rule for trust inference. AND and DT state that epistemic trust is closed under conjunction and disjunction, respectively. Liau (2003) rejects AND as a valid inference rule for trust, whereas Dastani et al. (2005)

proposes DT as an alternative. Finally, WT states that epistemic trust is closed under logical consequence, which is also rejected by Liau (2003). Recently, a normal modal logic for "trust in sincerity" was proposed by Leturc and Bonnet (2018) in which all the following rules are valid for "trust in sincerity" except for ST.

EQ	If $a \in T$ and $a \dashv\vdash b$, then $b \in T$.
ST	If $a \in T$, then $\neg a \in T$.
AND	If $a, b \in T$, then $a \wedge b \in T$.
DT	If $a, b \in T$, then $a \vee b \in T$.
WT	If $a \in T$ and $a \vdash b$, then $b \in T$.

We consider the following notions of constrained rationalizability.

Definition 10 (Constrained rationalizability)**.** Let $\mathbb{C} \subseteq \{\text{EQ}, \text{AND}, \text{DT}, \text{ST}, \text{WT}\}$. We say an acceptance function is $\mathbb{C}$-rationalizable if there exists a set T of formulas rationalizing α and T is closed under the constraints in $\mathbb{C}$.

It is natural to ask how these notions of constrained rationalizability affect the properties of acceptance functions. Since any constrained rationalizable acceptance functions are also rationalizable, they must satisfy the properties Exclusion – Groundedness. To find an adequate set of properties characterizing a $\mathbb{C}$-rationalizable function α, by Lemma 8 it suffices to show that $Fix(\alpha)$ is also closed under the constraints in $\mathbb{C}$.

Propostion 11. *Let $X \in \{\text{DT}, \text{WT}\}$, and T be a set of formulas closed under X. If α is rationalized by T, then $Fix(\alpha)$ is closed under X.*

Proof. Case DT. Suppose $\alpha(A) = Cn(T \cap Cn(A)) \cap A$ and T is closed under DT. Suppose $a \in \alpha(a)$ and $b \in \alpha(b)$. I.e.,

$$a \in Cn(T \cap Cn(a)) \cap \{a\} \text{ and } b \in Cn(T \cap Cn(b)) \cap \{b\}.$$

It follows that there are $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ in T such that $x_1 \wedge \cdots \wedge x_n \dashv\vdash a$ and $y_1 \wedge \cdots \wedge y_m \dashv\vdash b$. Note that $\bigwedge_{1 \le i \le n, 1 \le j \le m} x_i \vee y_j \dashv\vdash a \vee b$ and each $x_i \vee y_j$ is in T since T is closed under DT. Hence, $a \vee b \in \alpha(a \vee b)$.

Case WT. Suppose that $\alpha(A) = Cn(T \cap Cn(A)) \cap A$ and T is closed under WT. If $a \in \alpha(a) = Cn(T \cap Cn(a)) \cap \{a\}$, then there are $x_1, \ldots, x_n$ in T such that $x_1 \wedge \cdots \wedge x_n \dashv\vdash a$. Suppose $a \vdash b$, then $x_1 \wedge \cdots \wedge x_n \vdash b$. Therefore, $(x_1 \vee b) \wedge \cdots \wedge (x_n \vee b) \dashv\vdash b$. Note that for each i, $x_i \vee b \in T$ since T is closed under WT. Therefore, $b \in \alpha(b)$. □

Propostion 12. *Let $\mathbb{C} \subseteq \{\text{EQ}, \text{AND}, \text{DT}, \text{WT}\}$. Then the following statements are equivalent:*

(1) α is $\mathbb{C}$-rationalizable.

(2) α is $\mathbb{C} \setminus \{\text{EQ}, \text{AND}\}$-rationalizable.

(3) α satisfies Exclusion*,* Left strengthening*,* Right weakening*,* Monotony*,* Groundedness*, and $Fix(\alpha)$ is closed under all constraints in $\mathbb{C} \setminus \{\text{EQ}, \text{AND}\}$.*

Proof. The implication from (1) to (2) is obvious.
(2) ⇒ (3). This follows from Propositions 5 and 11.
(3) ⇒ (1). This follows from Lemma 8 and Prop. 7. □

If α is rationalized by a set T of formulas that is closed under ST, then $Fix(\alpha)$ need not be closed under ST. For example, let $\alpha(A) = Cn(\{p, q, \neg p, \neg q, \neg\neg p, \ldots\} \cap Cn(A)) \cap A$. Then $p \wedge q \in \alpha(p \wedge q)$. However, $\neg(p \wedge q) \notin \alpha(\neg(p \wedge q))$.

Next, we show that if ST is combined with one of AND, DT, and WT, then the preservation of constraints does hold.

Propostion 13. *Let $X \in \{\text{AND}, \text{DT}, \text{WT}\}$, and T be a* non-empty *set of formulas closed under X and* ST*. If α is rationalized by T, then $Fix(\alpha)$ is closed under X and* ST*.*

Proof. Suppose T is non-empty and closed under X and ST, and $\alpha(A) = Cn(T \cap Cn(A)) \cap A$.

Case X = AND. Since, by Proposition 7, $Fix(\alpha)$ is closed under AND, it suffices to show that $Fix(\alpha)$ is also closed under ST. If $a \in \alpha(a) = Cn(T \cap Cn(a)) \cap \{a\}$, then there are $x_1, \ldots, x_n$ in T such that $x_1 \wedge \cdots \wedge x_n \dashv\vdash a$. We distinguish three sub-cases: (1) $n = 0$. Then $a \dashv\vdash \top$. Since T is non-empty, let $b \in T$. By ST, $\neg b \in T$. Note that $b \wedge \neg b \dashv\vdash \neg a$. Hence, $\neg a \in \alpha(\neg a)$. (2) $n = 1$. Then $x_1 \dashv\vdash a$. By ST, $\neg x_1 \in T$. Thus, since $\neg x_1 \dashv\vdash \neg a$, $\neg a \in \alpha(\neg a)$. (3) $n > 1$. Since $x_1, \ldots, x_n \in T$, $x_1 \wedge \cdots \wedge x_n \in T$ by AND. Thus, $\neg(x_1 \wedge \cdots \wedge x_n) \in T$ by ST. Since $\neg(x_1 \wedge \cdots \wedge x_n) \dashv\vdash \neg a$, $\neg a \in \alpha(\neg a)$.

Case X = DT. Since, by Proposition 11, $Fix(\alpha)$ is closed under DT, it suffices to show that $Fix(\alpha)$ is also closed under ST. This can be shown similarly to the above.

Case X = WT. Since, by Proposition 11, $Fix(\alpha)$ is closed under WT, it suffices to show that $Fix(\alpha)$ is also closed under ST. Suppose $a \in \alpha(a) = Cn(T \cap Cn(a)) \cap \{a\}$. Then there are $x_1, \ldots, x_n$ in T such that $x_1 \wedge \cdots \wedge x_n \dashv\vdash a$. To show $\neg a \in \alpha(\neg a)$, we distinguish two sub-cases: (1) $n = 0$. This can be shown as above. (2) $n > 0$. Then $\neg x_1 \in T$ (by ST) and $\neg x_1 \vdash \neg a$. Note that $\neg x_1 \vee \neg a \dashv\vdash \neg a$ and $\neg x_1 \vee \neg a \in T$ (by WT). Thus, $\neg a \in \alpha(\neg a)$. □

Propostion 14. *Let $\mathbb{C}$ be a set of constraints such that* ST *$\in \mathbb{C}$ and $\mathbb{C} \cap \{\text{AND}, \text{DT}, \text{WT}\} \neq \emptyset$. Then the following statements are equivalent:*

(1) α is $\mathbb{C}$-rationalizable.

(2) $\alpha(A) = Cn(\emptyset) \cap A$ for all A, or α satisfies Exclusion*,* Left strengthening*,* Right weakening*,* Monotony*,* Groundedness*, and $Fix(\alpha)$ is closed under all constraints in $\mathbb{C}$.*

Proof. (1) ⇒ (2). Suppose α is rationalized by a set T that is closed under all properties in $\mathbb{C}$. If $\alpha(A) \neq Cn(\emptyset) \cap A$ for some A, then $T \neq \emptyset$. By Proposition 5, α satisfies the mentioned properties. Moreover, since $T \neq \emptyset$, by Propositions 13 and 11 it follows that $Fix(\alpha)$ is closed under all constraints in $\mathbb{C}$.

(2) ⇒ (1). If $\alpha(A) = Cn(\emptyset) \cap A$ for all A, then α is rationalized by $\emptyset$. Note that $\emptyset$ is closed under the constraints in $\mathbb{C}$. Otherwise, suppose that α satisfies the mentioned properties and $Fix(\alpha)$ is closed under all constraints in $\mathbb{C}$. Note that, by Lemma 8, α is rationalized by $Fix(\alpha)$. Thus, α is $\mathbb{C}$-rationalizable. □

Remark 4. The representation results for ST- and $\{\text{ST}, \text{EQ}\}$-rationalizability remain to be explored. It seems difficult to

find natural properties of acceptance functions that characterize the two types of rationalizability. We leave this for future work. Another underexplored topic is the equivalence between various notions of constrained rationalizability, which may help identify the redundancy of certain inference rules for epistemic trust.

4 The Black-list Approach

In this section, we turn our attention to the black-list acceptance mode. As mentioned before, it means that the agent accepts anything except those that are distrusted. As a dual to the notion of "rationalizability", we introduce the notion of "b-rationalizability" to formalize the black-list acceptance mode of the agent. We propose properties of acceptance functions that completely characterize the class of all b-rationalizable acceptance functions. Furthermore, we also consider the notions of constrained b-rationalizability.

Let a set D of distrusted statements be given (i.e., $a \in D$ means that "the agent distrusts the source on a"). The naive way to formalize the black-list acceptance mode is to define $\alpha(A) = Cn(Cn(A) \setminus D) \cap A$. However, this does not work, as distrusted statements may still be in the output of α. For example, if $D = \{p\}$ and $A = \{p, q\}$, then $\alpha(A) = \{p, q\}$. We note that this problem has long been known in the belief revision literature (van Ditmarsch, van der Hoek, and Kooi 2008).

To deal with this difficulty, we employ the idea of "remainders" from the study on belief revision (Alchourrón, Gärdenfors, and Makinson 1985).

Definition 15. Given two sets A, B of formulas, a set $X \subseteq A$ is called a *maximal subset of A that fails to imply B* if

- $Cn(X) \cap B = \emptyset$, and
- for any Y with $X \subsetneq Y \subseteq A$, $Cn(Y) \cap B \neq \emptyset$.

$A \perp B$ (the remainder set) denotes the set of all maximal subsets of A that fail to imply B.

The following fact about the remainder set is useful for later proofs.

Lemma 16. *For any sets A, D of formulas and $X \in Cn(A) \perp D$, $Cn(X) = X$. Thus, $Cn(\bigcap(Cn(A) \perp D)) = \bigcap(Cn(A) \perp D)$.*

Now we are ready to present the notion of "b-rationalizability".

Definition 17. A set of formulas D *b-rationalizes* an acceptance function α if $\alpha(A) = \bigcap(Cn(A) \perp D) \cap A$. α is b-rationalizable if there is a set D of formulas such that D b-rationalizes α.

To illustrate the definition, consider an example:

Example 5 (Example 2 continued)**.** Suppose we only have two atoms p and q. Let $D = \{p \wedge q\}$ and β be an acceptance function b-rationalized by D.

- Since $Cn(p) \perp D = \{Cn(p)\}$, $\beta(p) = \{p\}$. By symmetry, $\beta(q) = \{q\}$.
- Note that $Cn(p, q) \perp D = \{Cn(p), Cn(q), Cn(p \leftrightarrow q)\}$. Thus, $\beta(p, q) = \emptyset$.

The example above shows that Monotony fails for b-rationalizable acceptance functions. Nevertheless, we can show that b-rationalizable acceptance functions satisfy all remaining properties from Section 3. Below we show only the cases of Right weakening and Left strengthening. The other cases will become clear later.

Propostion 18. *Let α be b-rationalizable. Then α satisfies* Right weakening *and* Left strengthening.

Proof. Suppose α is b-rationalized by D.

Right weakening. We need to show that $Cn(\alpha(A)) \cap A \subseteq \alpha(A)$. Note that $Cn(\alpha(A)) \subseteq Cn(\bigcap(Cn(A) \perp D)) \subseteq \bigcap(Cn(A) \perp D)$ (by Lemma 16). Hence, $Cn(\alpha(A)) \cap A \subseteq \bigcap(Cn(A) \perp D) \cap A = \alpha(A)$.

Left strengthening. Note that $\alpha(Cn(A)) \cap A = \bigcap(Cn(A) \perp D) \cap Cn(A) \cap A = \alpha(A)$. □

Our next goal is to find properties that completely characterize all b-rationalizable acceptance functions. Below we list some new properties of acceptance functions.

T	$\top \in \alpha(\top)$.
Weakening	If $B \subseteq Cn(\alpha(A))$, then $B \subseteq \alpha(B)$.
Union	If $B \subseteq Cn(A)$, then $\alpha(B) \cup \alpha(A) \subseteq \alpha(\alpha(B) \cup \alpha(A))$.
Groundedness$^-$	If $x \in A \setminus \alpha(A)$, then there is a finite $B \subseteq Cn(A)$ such that $\alpha(B) \cup \{x\} \not\subseteq \alpha(\alpha(B) \cup \{x\})$.

The property T is self-evident. Weakening states that if B follows from what are accepted in A, then B should be accepted when only itself is informed. This is because nothing distrusted can be inferred from B in this case. Note that Weakening is not a property satisfied by *rationalizable* acceptance functions, see Example 3. In addition, it can be seen that Weakening is a stronger form of Groundedness, see the next proposition.

Clearly, Union is a stronger form of Idempotence. Groundedness$^-$ says that if x is not accepted in A, then there must be a "reason" $\alpha(B)$ for the rejection of x, i.e., adding x to $\alpha(B)$ would make certain distrusted statements derivable. Groundedness$^-$ can roughly be seen as the converse of Union. In addition, one can show that Groundedness$^-$ follows from Monotony, see below.

Propostion 19. *The following hold for any acceptance function α with $\alpha(\emptyset) = \emptyset$:*

(1) if α satisfies Monotony*, then it satisfies* Groundedness$^-$.

(2) if α satisfies Union*, then it satisfies* Idempotence *and* R-Idempotence.

(3) if α satisfies Weakening*, then it satisfies* Groundedness.

Proof. We show only (1). Suppose α satisfies Monotony and $x \in A \setminus \alpha(A)$. To show α satisfies Groundedness$^-$, it suffices to show that $x \notin \alpha(x)$. Suppose not. Then by Monotony, it follows that $x \in \alpha(A)$. Contradiction! □

Next we verify the soundness of the four properties T – Groundedness$^-$ for b-rationalizable acceptance functions.

Propostion 20 (Soundness)**.** *Let α be a b-rationalizable acceptance function. Then α satisfies* Exclusion*,* T*,* Weakening*,* Union*, and* Groundedness$^-$.

Proof. Suppose α is b-rationalized by D. The case for Exclusion is trivial.

T. Note that $\alpha(\top) = \bigcap(Cn(\top) \perp D) \cap \{\top\}$. For each $X \in Cn(\top) \perp D$, by Lemma 16 it follows that $\top \in Cn(X) \subseteq X$. Hence, $\top \in \alpha(\top)$.

Weakening. Suppose $B \subseteq Cn(\alpha(A))$. We distinguish two cases: (1) $D \cap Cn(\emptyset) \neq \emptyset$. Then $Cn(B) \perp D = \emptyset$. Thus, $\alpha(B) = B$. (2) $Cn(\emptyset) \cap D = \emptyset$. Then $Cn(A) \perp D \neq \emptyset$. Since $B \subseteq Cn(\alpha(A))$, $B \subseteq Cn(\bigcap(Cn(A) \perp D)) \subseteq \bigcap(Cn(A) \perp D)$ (by Lemma 16). Thus, there must be $X \in Cn(A) \perp D$ such that $B \subseteq X$. Since $Cn(X) \cap D = \emptyset$, $Cn(B) \cap D = \emptyset$. Thus, $B \subseteq \alpha(B)$.

Union. Suppose $B \subseteq Cn(A)$. The case $D \cap Cn(\emptyset) \neq \emptyset$ is trivial since $\alpha(C) = C$ for all C in this case. Suppose $D \cap Cn(\emptyset) = \emptyset$. To show that $\alpha(B) \cup \alpha(A) \subseteq \alpha(\alpha(B) \cup \alpha(A))$, it suffices to show the following claim:

$$Cn(\alpha(B) \cup \alpha(A)) \cap D = \emptyset.$$

Note that $\alpha(B) \subseteq B \subseteq Cn(A)$ and $Cn(\alpha(B)) \cap D = \emptyset$. Hence, by the Lindenbaum lemma, there is $X \in (Cn(A) \perp D)$ such that $\alpha(B) \subseteq X$. Since $\alpha(A) \subseteq X$, $\alpha(B) \cup \alpha(A) \subseteq X$. Since $Cn(X) \cap D = \emptyset$, $Cn(\alpha(B) \cup \alpha(A)) \cap D = \emptyset$.

Groundedness$^-$. Suppose $x \in A$ and $x \notin \alpha(A) = \bigcap(Cn(A) \perp D) \cap A$. Then there must exist $X \in Cn(A) \perp D$ such that $x \notin X$. Note that $\alpha(X) = X$ and, by the maximality of X, $Cn(X \cup \{x\}) \cap D \neq \emptyset$. Let $y \in Cn(X \cup \{x\}) \cap D$. By the compactness of Cn, there is a finite $B \subseteq X$ such that $B \cup \{x\} \vdash y$. Thus, $B \cup \{x\} \not\subseteq \alpha(B \cup \{x\})$. Note that, since $Cn(B) \subseteq Cn(X)$, $Cn(B) \cap D = \emptyset$. Thus, $B = \alpha(B)$. Therefore, $\alpha(B) \cup \{x\} \not\subseteq \alpha(\alpha(B) \cup \{x\})$. □

Remark 5. From the above result, it follows that the acceptance function α from Example 1 is not b-rationalizable, because it does not satisfy Weakening.

Given Proposition 19, the following corollary of Proposition 20 follows immediately.

Corollary 21. *Let α be b-rationalizable. Then α satisfies* R-Idempotence*,* Idempotence *and* Groundedness*.*

Next we will show that the properties Exclusion and T – Groundedness$^-$ are sufficient to characterize all b-rationalizable acceptance functions. To this end, the next proposition is very useful. Given an acceptance function α, let $Dis(\alpha) = \{a \in \mathcal{L} \mid a \notin \alpha(A) \text{ for all } A\}$.

Propostion 22. *An acceptance function α is b-rationalizable iff $Dis(\alpha)$ b-rationalizes α.*

Proof. We show only the only-if part. Suppose α is b-rationalized by a set D of formulas. We distinguish 2 cases:

(1) $D \cap Cn(\emptyset) \neq \emptyset$. Then $\alpha(A) = A$ for all A. Thus, $Dis(\alpha) = \emptyset$. Note that $\bigcap(Cn(A) \perp \emptyset) \cap A = A$ for all A. Therefore, $Dis(\alpha)$ b-rationalizes α.

(2) $D \cap Cn(\emptyset) = \emptyset$. To show that $\alpha(A) = \bigcap(Cn(A) \perp Dis(\alpha)) \cap A$, it suffices to show that for all $X \subseteq Cn(A)$, $Cn(X) \cap D = \emptyset$ if and only if $Cn(X) \cap Dis(\alpha) = \emptyset$.

From right to left. It suffices to show that $D \subseteq Dis(\alpha)$. Let $a \in D$. For any input B, $\alpha(B) = \bigcap(Cn(B) \perp D) \cap B$. Since $Cn(\emptyset) \cap D = \emptyset$, by the Lindenbaum lemma, there must be an $X \in Cn(B) \perp D$. Note that $a \notin X$. Thus, $a \notin \alpha(B)$. Since B is arbitrary, $a \in Dis(\alpha)$.

From left to right. Suppose $Cn(X) \cap D = \emptyset$. Thus, $\alpha(Cn(X)) = \bigcap(Cn(X) \perp D) \cap Cn(X) = Cn(X)$. For each $a \in Cn(X)$, we have $a \in \alpha(Cn(X))$. Thus, $a \notin Dis(\alpha)$. Hence, $Cn(X) \cap Dis(\alpha) = \emptyset$. □

Note that $Dis(\alpha)$ is the inclusion-maximal set that b-rationalizes α, given that α is b-rationalizable. In addition, $Dis(\alpha)$ enjoys the following properties. The first states that $Dis(\alpha)$ is the complement of $Fix(\alpha)$, and the second states that $Dis(\alpha)$ is closed under strengthening.

Propostion 23. *If α is a b-rationalizable acceptance function, then the following holds:*

(1) $Dis(\alpha) = \mathcal{L} \setminus Fix(\alpha)$.

(2) If $a \in Dis(\alpha)$ and $b \vdash a$, then $b \in Dis(\alpha)$.

Proof. (1). The inclusion $\subseteq$ is trivial. For the converse, if $a \notin Dis(\alpha)$, then $a \in \alpha(A)$ for some A. Since α satisfies Weakening (Proposition 20), $a \in \alpha(a)$. Thus, $a \in Fix(\alpha)$.

(2). Suppose $b \vdash a$. If $b \notin Dis(\alpha)$, then $b \in \alpha(A)$ for some A. Since α satisfies Weakening (Proposition 20), $a \in \alpha(a)$. Thus, $a \notin Dis(\alpha)$. □

Now we are ready to present the representation theorem for b-rationalizability.

Lemma 24. *Let α satisfy* Exclusion*,* T*,* Weakening*,* Union*, and* Groundedness$^-$*. Then $\alpha(A) = \bigcap(Cn(A) \perp Dis(\alpha)) \cap A$ for all A.*

Proof. Suppose α satisfies the above properties. By T and Weakening, for all $a \in Dis(\alpha)$, $a \not\vdash \top$. We need to show that $\alpha(A) = \bigcap(Cn(A) \perp Dis(\alpha)) \cap A$.

The inclusion $\subseteq$. By Exclusion, it suffices to show that $\alpha(A) \subseteq X$ for all $X \in Cn(A) \perp Dis(\alpha)$. We first show the following claim:

Claim. $Cn(X \cup \alpha(A)) \cap Dis(\alpha) = \emptyset$.

Proof of claim. For each $y \in Cn(X \cup \alpha(A))$, by the compactness of Cn, there must be $x_1, \ldots, x_n \in X$ such that $\alpha(A) \cup \{x_1, \ldots, x_n\} \vdash y$. Note that $x_1 \wedge \cdots \wedge x_n \in Cn(X)$ and $Cn(X) \cap Dis(\alpha) = \emptyset$. Thus, $x_1 \wedge \cdots \wedge x_n \notin Dis(\alpha)$. By definition, there is B such that $x_1 \wedge \cdots \wedge x_n \in \alpha(B)$. Therefore, by Weakening, $x_1 \wedge \cdots \wedge x_n \in \alpha(x_1 \wedge \cdots \wedge x_n)$. Note that, since $x_1 \wedge \cdots \wedge x_n \in X \subseteq Cn(A)$, $\alpha(x_1 \wedge \cdots \wedge x_n) \cup \alpha(A) \subseteq \alpha(\alpha(x_1 \wedge \cdots \wedge x_n) \cup \alpha(A))$ by Union. Therefore, $\{x_1 \wedge \cdots \wedge x_n\} \cup \alpha(A) \subseteq \alpha(\{x_1 \wedge \cdots \wedge x_n\} \cup \alpha(A))$. By Weakening, it follows that $y \in \alpha(y)$. Thus, $y \notin Dis(\alpha)$. □

Since $Cn(X \cup \alpha(A)) \cap Dis(\alpha) = \emptyset$ and X is a maximal subset of $Cn(A)$ that fails to imply $Dis(\alpha)$, $\alpha(A) \subseteq X$.

The inclusion $\supseteq$. Let $x \notin \alpha(A)$. We show that $x \notin \bigcap(\overline{Cn(A) \perp Dis(\alpha)}) \cap A$. The case $x \notin A$ is trivial. Suppose $x \in A \setminus \alpha(A)$. It suffices to show that

$x \notin \bigcap(Cn(A) \perp Dis(\alpha))$. By Groundedness$^-$, there is a finite $B \subseteq Cn(A)$ such that $\alpha(B) \cup \{x\} \not\subseteq \alpha(\alpha(B) \cup \{x\})$. Note that $\alpha(B) \cap Dis(\alpha) = \emptyset$. Hence, there must be $X \in Cn(A) \perp Dis(\alpha)$ such that $\alpha(B) \subseteq X$. We show that $x \notin X$:

Suppose, towards a contradiction, $x \in X$. Note that $Cn(X) \cap Dis(\alpha) = \emptyset$. Since $\bigwedge(\alpha(B) \cup \{x\}) \in Cn(X)$, $\bigwedge(\alpha(B) \cup \{x\}) \notin Dis(\alpha)$, i.e., $\bigwedge(\alpha(B) \cup \{x\}) \in \alpha(C)$ for some C. Thus, $\alpha(B) \cup \{x\} \subseteq Cn(\alpha(C))$. By Weakening, $\alpha(B) \cup \{x\} \subseteq \alpha(\alpha(B) \cup \{x\})$. Contradiction! □

Theorem 2 (Representation)**.** *Let α be any acceptance function satisfying* Exclusion*,* T*,* Weakening*,* Union*, and* Groundedness$^-$*. Then α is b-rationalizable.*

4.1 Constraints on D

As in the white-list approach, we could consider different constraints on the set D of formulas to b-rationalize an acceptance function. In this subsection, we investigate the same constraints as in Subsection 3.1, i.e., EQ, ST, AND, DT, and WT.

Definition 25. Let $\mathbb{C} \subseteq \{\text{EQ}, \text{AND}, \text{DT}, \text{ST}, \text{WT}\}$. We say an acceptance function is $\mathbb{C}$-b-rationalizable if there is a set D of formulas b-rationalizing α and D is closed under all constraints in $\mathbb{C}$.

The first result in this subsection states that some (combinations of) closure properties on D would trivialize the acceptance function.

Propostion 26. *Let $\mathbb{C} \subseteq \{\text{EQ}, \text{ST}, \text{AND}, \text{DT}, \text{WT}\}$ and α be $\mathbb{C}$-b-rationalizable. Then the following holds:*

(1) If WT $\in \mathbb{C}$*, then $\alpha(A) = A$ for all A.*

(2) If $\{\text{DT}, \text{ST}\} \subseteq \mathbb{C}$, then $\alpha(A) = A$ for all A.

(3) If $\{\text{AND}, \text{ST}\} \subseteq \mathbb{C}$, then $\alpha(A) = A$ for all A.

Proof. (1) Suppose α is b-rationalized by D and D is closed under all constraints in $\mathbb{C}$. We distinguish two cases: (1) $D = \emptyset$. It is easy to see that $\alpha(A) = A$ for all A. (2) $D \neq \emptyset$. Since D is closed under WT, $\top \in D$. Thus, for all A, $Cn(A) \perp D = \emptyset$. Thus, $\alpha(A) = A$.

(2) and (3) can be shown similarly. □

Next, we consider the properties of constrained b-rationalizable acceptance functions. It is obvious that they enjoy all the properties of b-rationalizable acceptance functions, i.e., Exclusion, T, Weakening, Union, Groundedness$^-$. To find a complete set of properties characterizing $\mathbb{C}$-b-rationalizable acceptance functions, by Lemma 24 it suffices to show that $Dis(\alpha)$ is also closed under the constraints in $\mathbb{C}$. Next, we show that this holds for DT.

Propostion 27. *If α is* DT*-b-rationalizable, then $Dis(\alpha)$ is closed under* DT*.*

Proof. Suppose $\alpha(A) = \bigcap(Cn(A) \perp D) \cap A$ and D is closed under DT. Suppose $a, b \in Dis(\alpha)$. Then, since $Dis(\alpha) = \mathcal{L} \setminus Fix(\alpha)$ (Proposition 23(1)), $a \notin \alpha(a)$ and $b \notin \alpha(b)$. Hence, $Cn(a) \cap D \neq \emptyset$ and $Cn(b) \cap D \neq \emptyset$. Let $x \in Cn(a) \cap D$ and $y \in Cn(b) \cap D$. Note that $x \vee y \in D$ and $x \vee y \in Cn(a \vee b)$. Therefore, $Cn(a \vee b) \cap D \neq \emptyset$. Therefore, $a \vee b \notin \alpha(a \vee b)$. Thus, $a \vee b \in Dis(\alpha)$. □

Propostion 28. *Let $\mathbb{C} \subseteq \{\text{EQ}, \text{AND}, \text{DT}\}$. Then the following statements are equivalent:*

(1) α is $\mathbb{C}$-b-rationalizable.

(2) α is $\mathbb{C} \setminus \{\text{EQ}, \text{AND}\}$-b-rationalizable.

(3) α satisfies Exclusion*,* T*,* Weakening*,* Union*,* Groundedness$^-$*, and $Dis(\alpha)$ is closed under all constraints in $\mathbb{C} \setminus \{\text{EQ}, \text{AND}\}$.*

Proof. The implication from (1) to (2) is trivial. The implication from (2) to (3) follows from Proposition 27.

(3) $\Rightarrow$ (1). Since α satisfies the 5 properties, by Lemma 24 it follows that α is b-rationalized by $Dis(\alpha)$. Note that $Dis(\alpha)$ is closed under EQ and AND by Proposition 23(2). Hence, $Dis(\alpha)$ is closed under all constraints in $\mathbb{C}$. Therefore, α is $\mathbb{C}$-b-rationalizable. □

5 Related Work

Our work connects three areas: the logic of epistemic trust, abstract choice theory, and belief revision.

The logic of epistemic trust. Logical tools can be applied to study inferences about epistemic trust and its interaction with related notions such as belief and information. One of the earliest studies in the logic of epistemic trust is (Liau 2003), which proposes a family of modal logics to reason about these three concepts. A central axiom in Liau's logical systems states that if an agent is informed of a statement and trusts it, then she will also believe it; this is also the underlying idea of our white-list approach. Liau's work was followed by other logicians (e.g., (Dastani et al. 2005; Booth and Hunter 2018; Leturc and Bonnet 2018; Jiang and Naumov 2022; Li, Van der Torre, and Yu 2026)), who applied additional techniques such as belief revision and input output logic. However, as we mentioned before, these authors disagree on the appropriate inference rules for epistemic trust.

Existing logical studies of epistemic trust typically assume that an agent's epistemic trust is given and focus on how trust affects the acceptance of information. In reality, however, epistemic trust and distrust toward others may be private, and only acceptance and rejection behaviors are observable. We ask the reverse question: can an agent's epistemic trust and distrust be revealed from her acceptance behavior?

Abstract choice theory. Similar questions have been asked in economics; for example, can we reveal an agent's preference relation from her choices (Arrow 1951; Uzawa 1956)? This is studied in the area of economics called choice theory. In abstract choice theory (Chambers and Echenique 2016), given a set X of objects that can possibly be chosen and a collection Σ of subsets of X, a choice function is a mapping $c : \Sigma \to 2^X$ such that $c(B) \subseteq B$ for each $B \in \Sigma$. The central question is whether a given choice function c can be *rationalized* by a preference relation $\succeq \subseteq X \times X$, in the sense that $c(B)$ picks out exactly the best elements of B according to $\succeq$. Obviously, not all choice functions can be rationalized and, if $\succeq$ is subject to extra constraints (such as transitivity), then the class of rationalizable choice functions changes. The main objective of abstract choice theory

	Exclusion	Right weakening	Left strengthening	Monotony	R-Idempotence	Groundedness
Rationalizability	✓(Prop. 5)	✓(Prop. 5)	✓(Prop. 5)	✓(Prop. 5)	✓(Prop. 4 & 5)	✓(Prop. 5)
B-rationalizability	✓(Prop. 20)	✓(Prop. 18)	✓(Prop. 18)	× (Ex. 5)	✓(Cor. 21)	✓(Cor. 21)

	T	Weakening	Union	Idempotence	Groundedness$^-$
Rationalizability	✓	× (Ex. 3)	× (Ex. 4)	× (Ex. 4)	✓(Prop. 19 & 5)
B-rationalizability	✓(Prop. 20)	✓(Prop. 20)	✓(Prop. 20)	✓(Cor. 21)	✓(Prop. 20)

Table 1: The comparison of the white- and black-list approaches, where ✓indicates that the given notion of rationalizability satisfies the given properties, and × otherwise. The justifications are given in the brackets. The fact that rationalizable acceptance functions satisfy T is obvious.

is to identify necessary and sufficient conditions for a choice function to be rationalizable.

Our acceptance functions are essentially choice functions over propositional formulas. We proposed different notions of rationalizability for acceptance functions. In the white-list approach, for example, an acceptance function α is rationalizable if there exists a set T of trusted statements such that $\alpha(A)$ picks out those elements of A that can be inferred from the trusted information contained in A. Our representation theorem provides necessary and sufficient conditions for an acceptance function to be rationalizable, and we also studied rationalizability under different constraints on epistemic trust.

Given the similarities between our work and abstract choice theory, it is natural to compare the properties of acceptance functions with those of choice functions. Due to space limitations, we leave this comparison for future work.

Belief revision. In the black-list approach, our definition of "b-rationalizability" is inspired by (full meet) belief contraction in the AGM theory (Alchourrón, Gärdenfors, and Makinson 1985). The main idea is the "remainder set", which is also used in other non-monotonic reasoning formalisms, such as input output logic (Makinson and van der Torre 2001).

In belief contraction, given a set K of formulas representing one's beliefs, the question is how to obtain a new set of beliefs if one belief a is to be given up. Since a set of beliefs is required to be closed under logical consequences, simply removing a from K will not work. To deal with this, we need to consider the maximal subsets of K that do not imply a, i.e., the remainders. The full meet belief contraction operation $\ominus$ predicts that the new belief set is obtained by taking the intersection of all remainders, i.e., $K \ominus a = \bigcap(K \perp \{a\})$.

The full meet belief contraction $\ominus$ is a two-place function. One can define a unary function from $\ominus$ by setting $\ominus_a(K) = K \ominus a = \bigcap(K \perp \{a\})$. As pointed out by Makinson (2005, p. 144), this function can hardly be called a form of inference, because the inclusion fails. However, it makes perfect sense if we view it as a choice function (or, as we call it, an acceptance function). In fact, if $K = Cn(K)$, it is not hard to see that $\ominus_a(K) = \alpha(K)$ where α is an acceptance function b-rationalized by $\{a\}$. Thus, from the point of view of belief contraction, the acceptance function α constantly contracts the statement a from any incoming information K. The main difference here is that we assume the formula a is unknown and to be revealed, whereas in belief contraction, a is given. [2]

6 Conclusion and Future Work

In this paper, inspired by revealed preference theory in economics, we study the problem of whether an agent's hidden epistemic (dis)trust can be revealed from her observable acceptance behavior. We considered two acceptance modes: in the white-list mode, the agent accepts only what is trusted, whereas in the black-list mode, she accepts everything except what is distrusted. For each mode, we introduced formal notions of "rationalizability" for acceptance functions and provided simple postulates that characterize them. The comparison results are summarized in Table 1. The main difference is as follows: while the white-list mode generates monotonic acceptance functions, those generated by the black-list mode are generally non-monotonic and satisfy the additional properties Weakening and Union. Nevertheless, in both modes the acceptance functions satisfy basic properties such as Left strengthening and Right weakening.

Another observation is that, in each mode, rationalizable acceptance functions admit canonical rationalizers. For the white-list mode, the canonical rationalizer is the set of fixed points of the acceptance function, whereas for the black-list mode, it is the complement of the fixed point set.

For future work, there are several unresolved technical issues, including representation results for ST- and $\{\mathrm{ST}, \mathrm{EQ}\}$-rationalizability, as well as for some notions of constrained b-rationalizability. The equivalence between these notions of rationalizability also remains to be established. We can also consider restrictions on the domain of acceptance functions, which may lead to substantially different characterization results.

Another promising direction is to consider different degrees of (dis)trust, since an agent may trust another person's judgment of p more than that of q. This becomes even more interesting if the agent distrusts the judgment of $p \wedge q$, as in Example 5. In that case, if p and q are reported, then rather than rejecting both of them, the agent may still be inclined

[2] In the belief revision literature, there also exists research on how to find the revision formula if we only know the belief sets before and after the revision, see (Schwind et al. 2019; Booth and Hunter 2018). But they focus mainly on the algorithmic aspects.

to accept p. To model this, we may need selection functions on the remainder set, as in the belief revision literature.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work is supported by the Luxembourg National Research Fund (FNR) through the following projects: The Epistemology of AI Systems (EAI) (C22/SC/17111440), DJ4ME – A DJ for Machine Ethics: the Dialogue Jiminy (O24/18989918/DJ4ME), Logical Methods for Deontic Explanations (LoDEx) (INTER/DFG/23/17415164/LoDEx), and Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN) (C24/19003061/SERAFIN). This work is also supported by the project Norm-First & HPC-Verified Interactive Agents for Child-Facing Applications - AI4Kids (19842028). This HPC BRIDGES project benefits from shared financial support by the Ministry of Economy and the Luxembourg National Research Fund (FNR) (19842028).

AI Declaration

The authors have not employed any Generative AI tools.

References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic* 50(2):510–530.

Arrow, K. J. 1951. *Social choice and individual values*. Yale University Press.

Booth, R., and Hunter, A. 2018. Trust as a precursor to belief revision. *Journal of Artificial Intelligence Research* 61:699–722.

Chambers, C. P., and Echenique, F. 2016. *Revealed Preference Theory*. Econometric Society Monographs. Cambridge University Press.

Dastani, M.; Herzig, A.; Hulstijn, J.; and van der Torre, L. 2005. Inferring trust. In Leite, J., and Torroni, P., eds., *Computational Logic in Multi-Agent Systems*, 144–160. Berlin, Heidelberg: Springer Berlin Heidelberg.

Jiang, J., and Naumov, P. 2022. In data we trust: The logic of trust-based beliefs. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2683–2689. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Leturc, C., and Bonnet, G. 2018. A normal modal logic for trust in the sincerity. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '18, 175–183. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Li, X.; Van der Torre, L.; and Yu, L. 2026. A logical analysis of an information filtering architecture based on epistemic trust inference. *Proceedings of the AAAI Conference on Artificial Intelligence* 40(23):19259–19266.

Liau, C.-J. 2003. Belief, information acquisition, and trust in multi-agent systems – a modal logic formulation. *Artificial Intelligence* 149(1):31–60.

Makinson, D., and van der Torre, L. 2001. Constraints for input/output logics. *Journal of Philosophical Logic* 30(2):155–185.

Makinson, D. 2005. *Bridges from Classical to Nonmonotonic Logic*. King's College Publications.

Schwind, N.; Inoue, K.; Konieczny, S.; Lagniez, J.-M.; and Marquis, P. 2019. What Has Been Said? Identifying the Change Formula in a Belief Revision Scenario. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 1865–1871. Macao, China: International Joint Conferences on Artificial Intelligence Organization.

Sen, A. K. 1971. Choice functions and revealed preference. *The Review of Economic Studies* 38(3):307–317.

Uzawa, H. 1956. Note on preference and axioms of choice. *Annals of the Institute of Statistical Mathematics* 8(1):35–40.

van Ditmarsch, H.; van der Hoek, W.; and Kooi, B. 2008. *Dynamic Epistemic Logic*. Springer, Dordrecht.