Journal of Philosophical Logic (2023) 52:1637–1671
https://doi.org/10.1007/s10992-023-09721-8

Truth Meets Vagueness. Unifying the Semantic and the Soritical Paradoxes

Riccardo Bruni[1] · **Lorenzo Rossi**[2]

Received: 27 February 2022 / Accepted: 27 August 2023 / Published online: 9 October 2023

Abstract

Semantic and soritical paradoxes display remarkable family resemblances. For one thing, several non-classical logics have been independently applied to both kinds of paradoxes. For another, revenge paradoxes and higher-order vagueness—among the most serious problems targeting solutions to semantic and soritical paradoxes—exhibit a rather similar dynamics. Some authors have taken these facts to suggest that truth and vagueness require a unified logical framework, or perhaps that the truth predicate is itself vague. However, a common core of semantic and soritical paradoxes has not been identified yet, and no explanation of their relationships has been provided. Here we aim at filling this lacuna, in the framework of many-valued logics. We provide a unified diagnosis of semantic and soritical paradoxes, identifying their source in a general form of indiscernibility. We then develop our diagnosis into a theory of paradoxicality, which formalizes both semantic and soritical paradoxes as arguments involving specific instances of our generalized indiscernibility principle, and correctly predicts which logics can non-trivially solve them.

Introduction

The semantic paradoxes and the paradoxes of vagueness ('soritical paradoxes', after the Sorites Paradox) display remarkable family resemblances. For one thing, the same non-classical logics have been independently applied to both kinds of paradoxes—

✉ Riccardo Bruni
riccardo.bruni@unifi.it

Lorenzo Rossi
lo.rossi@unito.it

1 Dipartimento di Lettere e Filosofia (DILEF), Università degli Studi di Firenze, via della Pergola 60 50134 Firenze, Italy

2 Dipartimento di Filosofia e Scienze dell'Educazione (DFE) and Center for Logic, Language, and Cognition (LLC), Università degli Studi di Torino, via Sant'Ottavio 20, Palazzo Nuovo, 10124 Torino, Italy

including many-valued,[1] supervaluational,[2], non-transitive[3] logics. Indeed, the similarity extends to some classical theories of truth and vagueness, notably contextualist theories.[4] For another, both *revenge paradoxes* and *higher-order vagueness* (henceforth: HOV) paradoxes—among the most serious problems targeting solutions to semantic and soritical paradoxes—exhibit a rather similar dynamics [67, 70]. Revenge arguments aim at showing that a given solution to the semantic paradoxes cannot extend to further semantic notions, by employing variants of the 'standard' semantic paradoxes such as the Liar or Curry's Paradox. Similarly, the HOV paradoxes aim at showing that a given solution to the soritical paradoxes cannot extend to further vague predicates, by employing variants of the 'standard' Sorites.

These facts have been taken by some authors to suggest that truth and vagueness require a unified logical framework [12, 22, 72]. Some authors go further, and argue that truth is itself a vague or indeterminate concept [17, 52]. Importantly, however, there currently is no identification, much less a formal theory, of what the common features of semantic and soritical paradoxes exactly consist in. This is what we aim to do in this work. More specifically, we develop (i) a theory of paradoxicality, where semantic and soritical paradoxical arguments can be studied, compared, and then unified; (ii) a reduction of the truth- and vagueness-theoretic principles responsible for semantic and soritical paradoxes to a common underlying principle, a form of *indiscernibility*. More about (i) and (ii) in a moment.

For the sake of concreteness and space, we focus on treatments of paradoxes within a specific family of non-classical logics, that is 4 three-valued logics, which are commonly adopted for truth and vagueness. However, the unification and reduction of paradoxes we propose has a much wider scope. On the one hand, the theory of paradoxicality we develop can be easily generalized to any truth-functional semantics, thus covering several generalizations of our four target logics, and much more. On the other, the indiscernibility principle we isolate as the source of the paradoxes is not specific to any logic.

As anticipated, the paper has two main outcomes, i.e. (i) and (ii) above. As for (i), we give a formal analysis of paradoxes, which then produces their unification. In order to do so, we develop an enrichment of standard model-theoretic semantics – called *equational semantics* – where paradoxical arguments which cannot be represented in standard model-theoretic semantics can be fully formalized and studied. We then apply our semantics to the case of semantic, soritical, revenge, and HOV paradoxes in our target logics, showing that the target theories treat them in the same way: any such theory blocks semantic paradoxes just in case it blocks soritical paradoxes, and fails to block revenge paradoxes just in case it fails to block HOV paradoxes. As for (ii), we motivate and articulate a diagnosis of what generates both kinds of paradoxess, which we identify in a general form of *indiscernibility*. We then formally derive the truth- and vagueness-theoretic principles involved in semantic and soritical paradoxes from our indiscernibility principle, and interpret all the target paradoxical

[1] See e.g. [3, 24, 25, 38, 43, 49, 62, 64].

[2] See e.g. [9, 26, 49].

[3] See e.g. [10–12, 66, 73, 77, 79].

[4] See e.g. [30, 31, 36, 74].

reasonings as arguments involving instances of it. Let us emphasize that the goal of this paper is *not to develop a new solution* to the paradoxes. Having effectively unified both the functioning and the root of semantic and soritical paradoxes, we argue that this reduction allows us to assess existing theories of truth and vagueness, to correctly predict which logics can non-trivially solve the paradoxes, and to guide to the development of new, unified theories.

The paper has the ambition to speak to different types of readers, and presupposes minimal background. Sections 1-3 concisely present our target topics, while Sections 4-5 develop our unification and reduction of the paradoxes.

1 Paradoxes and Three-Valued Logics

In Section 1, we introduce a formal language that is expressive enough to formulate the principles that are at the root of semantic and soritical paradoxes, and four three-valued logics that have been employed to address them. Section 2 introduces theories of truth over those three-valued logics, and their treatment of semantic and revenge paradoxes. Section3 presents theories of vagueness built over the same logics, together with their treatment of soritical and HOV-paradoxes.

1.1 Languages and Models

We introduce a formal language that satisfies the minimal requirements that are needed to formulate both semantic and soritical paradoxes.

Definition 1.1 $\mathcal{L}_{\mathrm{t,v}}$ is a first-order language (including a propositional constant $\bot$ for 'absurdity') that satisfies the following requirements:

(i) $\mathcal{L}_{\mathrm{t,v}}$ includes a predicate Tr, and countably many predicates $\mathsf{P}_1, \mathsf{P}_2, \ldots, \mathsf{P}_n, \ldots$
(ii) For every P_i, $\mathcal{L}_{\mathrm{t,v}}$ includes one binary relation constant $\sim_{\mathsf{P}_i}$ and countably many individual constants $c_1^{\mathsf{P}_i}, c_2^{\mathsf{P}_i}, , \ldots, c_n^{\mathsf{P}_i}, \ldots$ (for simplicity, we omit the superscripts).
(iii) It is possible to define in $\mathcal{L}_{\mathrm{t,v}}$ a function $\ulcorner \urcorner$ s.t. for every $\mathcal{L}_{\mathrm{t,v}}$-formula φ, $\ulcorner\varphi\urcorner$ is a closed term.
(iv) There is at least one $\mathcal{L}_{\mathrm{t,v}}$-structure $\mathcal{M}$ with support M s.t. (a) M is countable, (b) $\mathcal{M}$ is *acceptable* in the sense of Moschovakis, (c) for every $a \in M$ there is an $\mathcal{L}_{\mathrm{t,v}}$-constant c_a.
(v) For every open $\mathcal{L}_{\mathrm{t,v}}$-formula $\varphi(x)$, there is an $\mathcal{L}_{\mathrm{t,v}}$-term t_φ s.t. $t_\varphi = \ulcorner\varphi(t_\varphi/x)\urcorner$ in the selected acceptable model, where $\varphi(t_\varphi/x)$ is the result of uniformly replacing every free occurrence of x with t_φ in φ.

Requirement (i) makes sure that the language $\mathcal{L}_{\mathrm{t,v}}$ features a predicate for 'truth' (Tr) and vague predicates ($\mathsf{P}_1, \mathsf{P}_2, \ldots, \mathsf{P}_n, \ldots$). Requirement (ii) ensures that for each vague predicate P there is a similarity relation ($\sim_{\mathsf{P}}$), and names of individuals that can be used in a soritical series. Requirements (iii)-(v) make sure that $\mathcal{L}_{\mathrm{t,v}}$ can be used to

formalize truth-predications. The truth predicate is a predicate and, as such, it applies to terms. However, the relevant applications of the truth predicate are to terms that *denote sentences*, as "'Charpentier was a great composer' is true". This sentence has the form $\mathsf{Tr}(t)$, where t is the term 'Charpentier was a great composer', i.e. a term denoting a sentence (a *name* for a sentence).[5] Therefore, requirement (iii) makes sure that, for every $\mathcal{L}_{\mathrm{t,v}}$-sentence φ, there is a term $\ulcorner\varphi\urcorner$ denoting it. $\ulcorner\varphi\urcorner$ can be understood as a name of φ, so that the function $\ulcorner\cdot\urcorner$ can be taken to work as quotation marks. Requirement (iv) is a technical requirement, and goes hand in hand with requirement (v): a model is *acceptable* if, essentially, it has an in-build, well-behaved coding mechanism—exactly of the kind used in defining $\ulcorner\varphi\urcorner$.[6] The requirement that $\mathcal{L}_{\mathrm{t,v}}$ has constants for each element of at least one (countable) acceptable model dispenses with the need to introduce variable assignments. Finally, requirement (v) makes sure that it is possible to formulate paradoxical sentences. Consider the open formula $\neg\mathsf{Tr}(x)$, i.e. 'x is not true'. By (v), there is a term, call it t_λ, that denotes $\ulcorner\neg\mathsf{Tr}(t_\lambda)\urcorner$ in the selected acceptable model. Let's use λ to abbreviate the sentence $\neg\mathsf{Tr}(t_\lambda)$. λ is a *Liar sentence* and can be informally interpreted as saying that t_λ is not true. But what is t_λ? Is a name of $\neg\mathsf{Tr}(t_\lambda)$, i.e. a name of λ itself. Therefore, there is a sense in which λ says of itself that it is not true.[7]

$$\mathrm{MDIAG}_\lambda \ \frac{\Gamma \vdash \lambda}{\Gamma \vdash \neg\mathsf{Tr}(t_\lambda)}$$

where $\vdash$ is whichever consequence relation we will be employing.[8]

Terms, closed terms, formulae, and closed formulae (i.e. sentences) of $\mathcal{L}_{\mathrm{t,v}}$ are defined as usual. We use s and t to range over $\mathcal{L}_{\mathrm{t,v}}$-terms, φ, ψ, and χ to range over $\mathcal{L}_{\mathrm{t,v}}$-formulae, and Γ and Δ to range over sets of $\mathcal{L}_{\mathrm{t,v}}$-formulae. We take $\neg$, $\wedge$, and $\forall$ as primitive, while $\vee$, $\rightarrow$, $\leftrightarrow$, and $\exists$ are defined in the usual way. Open terms and formulae will be explicitly indicated (as in $t(x)$ and $\varphi(x)$). '$\varphi \in \mathcal{L}_{\mathrm{t,v}}$' and '$\Gamma \subseteq \mathcal{L}_{\mathrm{t,v}}$' are abbreviations for 'φ is an $\mathcal{L}_{\mathrm{t,v}}$-sentence' and '$\Gamma$ is a set of $\mathcal{L}_{\mathrm{t,v}}$-sentences' respectively. 'S.t.' abbreviates 'such that'.

[5] Applications of the truth predicate to terms that do *not* denote sentences, as in 'the number zero is true', seem to be misapplications of the truth predicate. Similarly, we are not considering cases where 'true' is used as a synonym of 'authentic', as in 'a true friend', or 'a true Chagall'.

[6] For the definition, see [55], Ch. 4. Acceptability will ensure that the semantic construction in Section 2.1, due to [49], can actually be carried out.

[7] Requirement (v) requires $\mathcal{L}_{\mathrm{t,v}}$ to have a constant for the primitive recursive function of substitution, and theories formulated in $\mathcal{L}_{\mathrm{t,v}}$ to have defining equations for it. For details see [42, 60]. Requirement (v) enables us to employ the sentence-formation process just described—'strong diagonalization'—and employ it in inferences in any theory that we are going to consider. More explicitly, we are going to avail ourselves of a meta-rule of inference of the following kind (we exemplify it here with λ and $\neg\mathsf{Tr}(t_\lambda)$):

[8] This meta-inferential formulation of diagonalization is required, because the 'usual' form of diagonalization ('weak diagonalization') involving a biconditional is not available in some of the theories we consider (e.g. those based on the paracomplete logic K3: see [52], p. 111 and following), and the inferential form of diagonalization ($\lambda \vdash \neg\mathsf{Tr}(\ulcorner\lambda\urcorner)$ and $\neg\mathsf{Tr}(\ulcorner\lambda\urcorner) \vdash \lambda$ fails in theories based on non-reflexive logics.

1.2 Naïveté and the Semantic Paradoxes

The truth predicate is often argued to satisfy a property of *naïveté*, to the effect that, for any sentence φ, φ and $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ are in some sense equivalent. One way to spell out naïveté requires that all the instances of the following schema be validated:

(Tr- SCHEMA) $\varphi \leftrightarrow \mathsf{Tr}(\ulcorner\varphi\urcorner)$.

Alternatively, one can require naïve rules, to the effect that $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ can be always inferred from φ, and *vice versa*:

$$\text{Tr-INTRO}\ \frac{\Gamma \vdash \varphi}{\Gamma \vdash \mathsf{Tr}(\ulcorner\varphi\urcorner)} \qquad \frac{\Gamma \vdash \mathsf{Tr}(\ulcorner\varphi\urcorner)}{\Gamma \vdash \varphi}\ \text{Tr-ELIM}$$

where $\vdash$ is the consequence relation of the target theory of truth. Finally, the truth predicate might be required to obey an *inter-substitutivity* requirement, to the effect that φ and $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ are always intersubstitutable (in all non-opaque contexts). More precisely, it is required that from ψ one can always infer any formula ψ^{t} that results from ψ by replacing, possibly non-uniformly, a subformula φ of ψ with $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ or *vice versa*. Let's call ψ^{t} a *truth-theoretic substitution* of ψ.

A naïveté requirement for truth seems well-motivated by both linguistic reflection on the behavior of the truth predicate in natural languages, and the role of the truth predicate in formulating truth-conditions.[9] Moreover, naïveté seems required for the truth predicate to fulfill its expressive role in expressing agreement and disagreement.[10] What matters for us is that virtually all formulations of naïveté give rise to semantic paradoxes, over sufficiently strong logic and base theory.

Now, what is a semantic paradox? As far as we know, there is no standard definition in the literature. Typically, when semantic paradoxes are introduced (in a classroom, a paper, or a book), they are presented via representative examples, such as the Liar or Curry's Paradox, and no general definition is provided. Such examples take two main forms: proof-theoretical proofs of triviality – i.e. derivations of any sentence φ in a given formal systems that incorporates logical rules, the naïve truth-theoretical principles, and the definition of sentences such as λ, as in [56] –, or as model-theoretic proofs that use sentences such as λ to show that no classical interpretation of the language exists that is consistent with naïveté. Since our approach is mainly model-theoretic, we will only provide model-theoretic examples of paradoxes, and later model-theoretic characterizations. Here is a model-theoretic presentation (from [67]).

Example 1.2 (The Liar Paradox, as a model-theoretic non-existence proof) Suppose there is a classical valuation v s.t. $v(\varphi) = v(\mathsf{Tr}(\ulcorner\varphi\urcorner))$, for every $\varphi \in \mathcal{L}_{\mathrm{t,v}}$. Since v is a classical valuation, either $v(\lambda) = \mathbf{1}$ or $v(\lambda) = \mathbf{0}$.

– If $v(\lambda) = \mathbf{1}$, then $v(\neg\mathsf{Tr}(\ulcorner\lambda\urcorner)) = \mathbf{1}$ (by definition of λ), but also $v(\neg\lambda) = \mathbf{1}$ (by naïveté), which is absurd.

[9] A *locus classicus* is [14]. See [34] for the claim that truth-conditions for natural language sentences should be formulated via a self-applicable truth predicate.

[10] See [24] (Ch. 13), [61], and [57] .

– If $v(\lambda) = \mathbf{0}$, then $v(\neg \mathsf{Tr}(\ulcorner \lambda \urcorner)) = \mathbf{0}$ (by definition of λ), but also $v(\neg\lambda) = \mathbf{0}$ (by naïveté), which is also absurd.

Since both $v(\lambda) = \mathbf{1}$ and $v(\lambda) = \mathbf{0}$ lead to absurdity, there is no such valuation v.

1.3 Tolerance and the Soritical Paradoxes

Vague predicates (such as 'rich', 'tall', 'red', ...) are often argued to satisfy a property of *tolerance*. Let P be a vague predicate. Tolerance for P dictates that, if s is P and t is very similar to s as far as P is concerned (in symbols, $s \sim_{\mathsf{P}} t$), then t is P as well.

(TOLERANCE) $\forall x \forall y (\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y))$

As with naïveté, tolerance can also be formulated as an inference rule:

(TOLERANCE- INF) $\mathsf{P}(s_i) \wedge s_i \sim_{\mathsf{P}} s_j \vdash \mathsf{P}(s_j)$

or as a meta-inference rule:

$$\frac{\Gamma \vdash \mathsf{P}(s_i) \qquad \Delta \vdash s_i \sim_{\mathsf{P}} s_j}{\Gamma, \Delta \vdash \mathsf{P}(s_j)} \text{ TOLERANCE- META}$$

For simplicity, we will use TOLERANCE by default, keeping in mind that its inferential or meta-inferential formulation might be required in some of the theories we consider later.[11]

Just like naïveté, also tolerance for vague predicates is conceptually and linguistically well-motivated.[12] [78] goes so far as to argue that our understanding of vague predicates requires them to obey a tolerance principle. And, just like naïveté for truth, also tolerance gives rise to paradoxes (called 'soritical' from the adjective derived from *soros* ($\sigma\omega\rho\acute{o}\varsigma$) the ancient Greek word for 'heap', as the notion of heap was used to exemplify the paradox).

The soritical paradoxes have a more standard presentation than the semantic ones, and are typically presented as arguments displaying a structure of the following kind:

1. $\mathsf{P}(c_0)$	[Premiss 1]
2. $c_0 \sim_{\mathsf{P}} c_1$	[Premiss 2]
3. $\mathsf{P}(c_0) \wedge c_0 \sim_{\mathsf{P}} c_1$	[1, 2, $\wedge$-I]
4. $\forall x \forall y (\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y))$	[TOLERANCE]
5. $\mathsf{P}(c_0) \wedge c_0 \sim_{\mathsf{P}} c_1 \rightarrow \mathsf{P}(c_1)$	[3, $\forall$-E]
6. $\mathsf{P}(c_1)$	[3, 5, $\rightarrow$-E]

[11] For more details on the formalization of tolerance, see [13, 74]. Two immediate generalizations of tolerance involve (a) predicates of arbitrary arity:

$$\forall x_i, \ldots, \forall x_n \forall y_1, \ldots, \forall y_n (\mathsf{P}(x_i, \ldots, x_n) \wedge \langle x_i, \ldots, x_n \rangle \sim_{\mathsf{P}} \langle y_i, \ldots, y_n \rangle \rightarrow \mathsf{P}(y_i, \ldots, y_n)),$$

and (b) multi-dimensional vague predicates ([28, 29, 45]), i.e. predicates whose applicability is determined by several aspects of their meaning, such as 'intelligent'/'stupid', 'good'/'bad', 'successful'/'unsuccessful', etc. Generalizing tolerance in these two directions would bring us far afield from our main focus. However, it is clear that our results can be readily adapted to more complex versions of tolerance.

[12] See e.g. [8, 15, 16, 20, 41, 46, 54, 78]. For critical discussions, see e.g. [27, 37]. For empirical investigations into vague concepts, see e.g. [1, 18, 19, 40, 50, 75, 76].

7. ⋮ [reiterate the above passages, starting with 6]
   $\mathsf{P}(c_n)$

This reasoning is paradoxical because it seemingly shows that every individual has the property P, including those which clearly do not. Soritical paradoxes are not typically given model-theoretic presentations, but that is easy enough to do.

Example 1.3 (A soritical paradox for P, as model-theoretic proof). Suppose there is a classical valuation v s.t.:

$$\begin{aligned} &v(\mathsf{P}(c_0)) = 1 \\ &v(c_i \sim_{\mathsf{P}} c_{i+1}) = 1, \text{ for every } i \\ &v(\forall x \forall y(\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y))) = 1 \end{aligned}$$

Since v is classical, $v(\forall x \forall y(\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y))) = 1$ entails that $v(\mathsf{P}(c_0) \wedge c_0 \sim_{\mathsf{P}} c_1 \rightarrow \mathsf{P}(c_1)) = 1$, and since $v(\mathsf{P}(c_0)) = 1$ and $v(c_0 \sim_{\mathsf{P}} c_1) = 1$, also $v(\mathsf{P}(c_1))) = 1$. By induction, for every n, $v(\mathsf{P}(c_n)) = 1$.

1.4 Three-Valued Logics

In order to avoid semantic and soritical paradoxes, several authors have advocated the use of non-classical logics. For concreteness, here we focus on *three-valued* truth-functional logics.[13]

Definition 1.4 A *three-valued model* $\mathcal{M}$ is a pair $\langle M, f \rangle$, where M is a non-empty set and f is a multi-function from closed $\mathcal{L}_{\mathrm{t,v}}$-terms to M and from atomic $\mathcal{L}_{\mathrm{t,v}}$-sentences to the set $\{0, 1/2, 1\}$.

Using three-valued models, one can define *valuations* that extend the assignments of values in $\{0, 1/2, 1\}$ to logically complex formulae. A widely used valuation in three-valued logics is given by *strong Kleene* semantics [6, 47].

Definition 1.5 For every three-valued model $\mathcal{M} = \langle M, f \rangle$, the *strong Kleene valuation* induced by $\mathcal{M}$ is the function $v_{\mathcal{M}}$ from sentences to $\{0, 1/2, 1\}$ s.t.:

$$\begin{aligned} v_{\mathcal{M}}(R(t_0, \ldots, t_n)) &:= f(R(t_0, \ldots, t_n)) \\ v_{\mathcal{M}}(\neg\varphi) &:= 1 - v_{\mathcal{M}}(\varphi) \\ v_{\mathcal{M}}(\varphi \wedge \psi) &:= \min(v_{\mathcal{M}}(\varphi), v_{\mathcal{M}}(\psi)) \\ v_{\mathcal{M}}(\forall x \varphi(x)) &:= \inf\{v_{\mathcal{M}}(\varphi(t)) \in \{0, 1/2, 1\} \mid t \text{ is a closed } \mathcal{L}_{\mathrm{t,v}}\text{-term}\} \end{aligned}$$

[13] For semantic paradoxes, see e.g. [3, 6, 11, 24, 25, 38, 43, 49, 58, 62, 64, 66]. For soritical paradoxes, see e.g. [5, 16, 26, 35, 39, 51, 69]. [10, 12] present a simultaneous semantics for truth and vagueness.

The clauses of a strong Kleene valuation are just the classical valuation, generalized to $\{0, 1/2, 1\}$. Definitions 1.4 and 1.5 provide a semantics for $\mathcal{L}_{t,v}$ but not yet a logic. Using many-valued valuations, several notions of logical consequence are definable. We now present four logics that can be defined using strong Kleene semantics, following [10].

Definition 1.6 For every $\Gamma \subseteq \mathcal{L}_{t,v}$, a valuation e makes Γ *strictly true* (s-*true*) if, for every $\varphi \in \Gamma$, $v(\varphi) = 1$, and e makes Γ *tolerantly true* (t-*true*) if for every $\varphi \in \Gamma$, $v(\varphi) \geq 1/2$.

Definition 1.7 ss, tt, st, and ts

- Strict-strict logic. Γ ss-entails φ (in symbols $\Gamma \models_{ss} \varphi$) if for every three-valued model $\mathcal{M}$, every $v_{\mathcal{M}}$ induced by $\mathcal{M}$ that makes all the sentences in Γ s-true, makes φ s-true.
- Tolerant-tolerant logic. Γ tt-entails φ ($\Gamma \models_{tt} \varphi$) if for every three-valued model $\mathcal{M}$, every $v_{\mathcal{M}}$ induced by $\mathcal{M}$ that makes all the sentences in Γ t-true, makes φ t-true.
- Tolerant-strict logic. Γ ts-entails φ ($\Gamma \models_{ts} \varphi$) if for every three-valued model $\mathcal{M}$, every $v_{\mathcal{M}}$ induced by $\mathcal{M}$ that makes all the sentences in Γ t-true, makes φ s-true.
- Strict-tolerant logic. Γ st-entails φ ($\Gamma \models_{st} \varphi$) if for every partial three-valued $\mathcal{M}$, every $v_{\mathcal{M}}$ induced by $\mathcal{M}$ that makes all the sentences in Γ s-true, makes φ t-true.

ss is a *paracomplete* logic, that notably does not validate the introduction rules for negation and conditional: there are $\Gamma \cup \{\varphi, \psi\} \subseteq \mathcal{L}_{t,v}$ s.t.:

$\Gamma, \varphi \models_{ss} \bot$ but $\Gamma \not\models_{ss} \neg\varphi$
$\Gamma, \varphi \models_{ss} \psi$ but $\Gamma \not\models_{ss} \varphi \to \psi$

tt is a *paraconsistent* logic, which does not validate *ex falso quodlibet*, i.e. $\varphi \wedge \neg\varphi \not\models_{tt} \psi$, nor the elimination rules for negative connectives: there are $\Gamma \cup \{\varphi, \psi\} \subseteq \mathcal{L}_{t,v}$ s.t.:

$\Gamma \models_{tt} \varphi$ and $\Gamma \models_{tt} \neg\varphi$ but $\Gamma \not\models_{tt} \bot$
$\Gamma, \varphi, \varphi \to \psi \not\models_{tt} \psi$

ts is a *non-reflexive logic* ($\varphi \not\models_{ts} \varphi$) that does not validate any classical inference, but is closed under classically valid meta-inferences (e.g., the classically valid rules of a sequent calculus). Finally, st is a *non-transitive* logic, that validates all the classical laws and inferences, and does not validate some meta-inferences, including transitivity and *modus ponens* formulated as a meta-inference: there are $\Gamma \cup \{\varphi, \psi\} \subseteq \mathcal{L}_{t,v}$ s.t.:

$\Gamma \models_{st} \varphi$ and $\Delta \models_{st} \varphi \to \psi$ but $\Gamma, \Delta \not\models_{st} \psi$
$\Gamma \models_{st} \varphi$ and $\Delta, \varphi \models_{st} \psi$ but $\Gamma, \Delta \not\models_{st} \psi$

The logical differences between ss, tt, ts, and st determine also which form of naïveté and tolerance can be supported by these logics.[14]

[14] For more on ss, tt, ts, and st, see [59]. For the connections between ss and ts, and between tt and st, see [2].

2 Naïve Truth in Three-Valued Logics

2.1 Semantic Paradoxes in Three-Valued Logics

We now use strong Kleene semantics and the logics ss, tt, ts, and st to formulate theories of truth. In order to include a treatment of truth-predications, we move from a starting acceptable three-valued model $\mathcal{M} = \langle M, f \rangle$ to a triple $\langle M, f, S \rangle$, where S is the *extension* of the truth predicate, i.e. the elements of M to which Tr applies. The main model-theoretic technique to construct such an extension was articulated by [49], and it consists in building the extension of Tr in stages, indexed by ordinals. At stage 0, nothing is in the extension of Tr, so $\langle M, f, S^0 \rangle = \langle M, f, \varnothing \rangle$. At stage 1, Tr only applies to *atomic* sentences of the truth-free fragment $\mathcal{L}_{\mathrm{t,v}}$ which are satisfied by the starting three-valued model, and to negated atomic sentences of the truth-free fragment of $\mathcal{L}_{\mathrm{t,v}}$ which are not satisfied by the starting model. More formally:

$$\begin{aligned}\langle M, f, S^1 \rangle = \langle M, f, \{P(\bar{s}) \in \mathcal{L} \mid P \neq \mathsf{Tr} \text{ and } f(\bar{s}) \in f(P)\} \cup \\ \{\neg Q(\bar{t}) \in \mathcal{L} \mid Q \neq \mathsf{Tr} \text{ and } f(\bar{t}) \in f(\neg Q)\}\rangle\end{aligned}$$

For example, $s = s$ is in S^1: in every acceptable $\mathcal{M}$, the pair $\langle s, s \rangle$ is in the extension of the identity relation in $\mathcal{M}$. A successor stage of Kripke's construction is obtained by the application of a monotone operator on sets of (codes of) sentences. For any set of (codes of) sentences A, let Φ be the operator such that $\varphi \in \Phi(A)$ if:

φ is $\neg\neg\psi$ and $\psi \in A$, or
φ is $\psi \wedge \chi$, $\psi \in A$, and $\chi \in A$, or
φ is $\neg(\psi \wedge \chi)$ and ($\neg\psi \in A$ or $\neg\chi \in A$), or
φ is $\forall x \psi(x)$ and for every closed $\mathcal{L}$-term s, $\psi(s) \in A$, or
φ is $\neg\forall x \psi(x)$ and for some closed $\mathcal{L}$-term s, $\neg\psi(s) \in A$, or
φ is $\mathsf{Tr}(\ulcorner \psi \urcorner)$ and $\psi \in A$, or
φ is $\neg\mathsf{Tr}(\ulcorner \psi \urcorner)$ and $\neg\psi \in A$.

Φ takes (codes of) sentences in M as input, and outputs sentences that result from (i) combinations of such sentences that preserve value 1 in the strong Kleene valuations, (ii) truth-predications of sentences in M, and negated truth-predications of sentences whose negation is in M. Using Φ, we define the successor stage of Kripke's construction: $\langle M, f, S^{\alpha+1} \rangle = \langle M, f, S^\alpha \cup \Phi(S^\alpha) \rangle$. Finally, at limit stages, one takes unions. For λ a limit ordinal, put $\langle M, f, S^\lambda \rangle = \langle M, f, \bigcup_{\alpha<\lambda} S^\alpha \rangle$.

Kripke's construction of an extension for the truth predicate grows with ordinals. However, there are more ordinals than sets of $\mathcal{L}_{\mathrm{t,v}}$-sentences: ordinals are too many to form a set, while there are set-many sets of $\mathcal{L}_{\mathrm{t,v}}$-sentences. This means that there is a smallest (limit, and in our case countable) ordinal ζ s.t. at stage ζ, all the sentences that can be added to the extension of the truth predicate with this method have been added. In other words, the construction reaches a final stage, called a *fixed point*:

$$\Phi(S^\zeta) = S^\zeta \text{ and therefore } \langle M, f, S^\zeta \rangle = \langle M, f, S^{\zeta+1} \rangle$$

The process we described delivers the *least Kripke fixed point* S^{ζ}, and $\langle M, f, S^{\zeta}\rangle$ is the *least Kripke model* for $\mathcal{L}_{\mathrm{t,v}}$. More generally, every set A s.t. $\Phi(A) = A$ is a Kripke fixed point, and $\langle M, f, A\rangle$ is the corresponding Kripke model. For simplicity, we restrict ourself to *consistent* Kripke models, i.e. Kripke models $\langle M, f, A\rangle$ where no sentence ψ is in A together with its negation $\neg\psi$.

Let's now explicitly associate a valuation to a Kripke model.

Definition 2.1 For every Kripke model $\mathcal{M} = \langle M, f, S\rangle$ for $\mathcal{L}_{\mathrm{t,v}}$, the *Kripke (strong Kleene) valuation* induced by $\mathcal{M}$ is the function e from sentences to $\{0, 1/2, 1\}$ s.t.:

$$v_{\mathcal{M}}(\varphi) := \begin{cases} 1, & \text{if } \varphi \in S \\ 0, & \text{if } \neg\varphi \in S \\ 1/2, & \text{otherwise} \end{cases}$$

The following is immediate:

Lemma 2.2 *For every Kripke model $\mathcal{M}$, the valuation $v_{\mathcal{M}}$ is a strong Kleene valuation, and it validates a form of naïveté: for every $\varphi \in \mathcal{L}_{\mathrm{t,v}}$ and every truth-theoretic substitution φ^{t}, $v_{\mathcal{M}}(\varphi) = v_{\mathcal{M}}(\varphi^{\mathsf{t}})$.*

Let's call the above valuations 'Kripke-Kleene'. Finally, we associate theories of truth proper to the above models and valuations. We consider four such theories, corresponding to the four logics introduced above.

Definition 2.3 sst, ttt, stt, and tst

- Γ sst-entails φ ($\Gamma \models_{\mathsf{sst}} \varphi$) if for every Kripke model $\mathcal{M}$, if the Kripke-Kleene valuation $v_{\mathcal{M}}$ makes all the sentences in Γ s-true, it also makes φ s-true.
- Γ ttt-entails φ ($\Gamma \models_{\mathsf{ttt}} \varphi$) if for every Kripke model $\mathcal{M}$, if the Kripke-Kleene valuation $v_{\mathcal{M}}$ makes all the sentences in Γ t-true, it also makes φ t-true.
- Γ tst-entails φ ($\Gamma \models_{\mathsf{tst}} \varphi$) if for every Kripke model $\mathcal{M}$, if the Kripke-Kleene valuation $v_{\mathcal{M}}$ makes all the sentences in Γ t-true, it also makes φ s-true.
- Γ stt-entails φ ($\Gamma \models_{\mathsf{stt}} \varphi$) if for every Kripke model $\mathcal{M}$, if the Kripke-Kleene valuation $v_{\mathcal{M}}$ makes all the sentences in Γ s-true, it also makes φ s-true.

sst, ttt, tst, and stt share the same Kripke models, but their logical differences has an impact on the versions of naïveté they recover, as detailed in the next result (the proof is routine, and follows from Lemma 2.2).

Proposition 2.4

- *For every $\varphi \in \mathcal{L}_{\mathrm{t,v}}$, $\varphi \models_{\mathsf{sst}} \varphi^{\mathsf{t}}$, $\varphi \models_{\mathsf{ttt}} \varphi^{\mathsf{t}}$, and $\varphi \models_{\mathsf{stt}} \varphi^{\mathsf{t}}$.*
- *For some $\varphi \in \mathcal{L}_{\mathrm{t,v}}$, $\varphi \not\models_{\mathsf{tst}} \varphi^{\mathsf{t}}$ (e.g. letting φ be λ). However, for every $\Gamma \cup \{\varphi\} \subseteq \mathcal{L}_{\mathrm{t,v}}$:*

$$\frac{\Gamma \models_{\mathsf{tst}} \varphi}{\Gamma \models_{\mathsf{tst}} \varphi^{\mathsf{t}}}\ \mathrm{MSUB}_{\mathrm{Tr}}$$

- *Letting φ^{t} be $\mathsf{Tr}(\ulcorner\varphi\urcorner)$, it follows that the following naïve rules hold:*

$$\Gamma, \varphi \models_{\mathsf{sst}} \mathsf{Tr}(\ulcorner\varphi\urcorner); \quad \Gamma, \varphi \models_{\mathsf{ttt}} \mathsf{Tr}(\ulcorner\varphi\urcorner); \quad \Gamma, \varphi \models_{\mathsf{stt}} \mathsf{Tr}(\ulcorner\varphi\urcorner)$$
$$\Gamma, \mathsf{Tr}(\ulcorner\varphi\urcorner) \models_{\mathsf{sst}} \varphi; \quad \Gamma, \mathsf{Tr}(\ulcorner\varphi\urcorner) \models_{\mathsf{ttt}} \varphi; \quad \Gamma, \mathsf{Tr}(\ulcorner\varphi\urcorner) \models_{\mathsf{stt}} \varphi$$

$$\frac{\Gamma \models_{\mathsf{tst}} \varphi}{\Gamma \models_{\mathsf{tst}} \mathsf{Tr}(\ulcorner\varphi\urcorner)} \qquad \frac{\Gamma \models_{\mathsf{tst}} \mathsf{Tr}(\ulcorner\varphi\urcorner)}{\Gamma \models_{\mathsf{tst}} \varphi}$$

- *For every $\varphi \in \mathcal{L}_{\mathsf{t,v}}$:*

$$\models_{\mathsf{ttt}} \varphi \leftrightarrow \mathsf{Tr}(\ulcorner\varphi\urcorner) \qquad \models_{\mathsf{stt}} \varphi \leftrightarrow \mathsf{Tr}(\ulcorner\varphi\urcorner)$$

However:

$$\not\models_{\mathsf{sst}} \lambda \leftrightarrow \mathsf{Tr}(\ulcorner\lambda\urcorner) \qquad \not\models_{\mathsf{tst}} \lambda \leftrightarrow \mathsf{Tr}(\ulcorner\lambda\urcorner)$$

This completes the picture of three-valued logics applied to truth and semantic paradox. It is now easy to see how each of sst, ttt, stt, and tst blocks the paradoxical reasonings presented in Section 1.2.

2.2 Revenge Paradoxes in Three-Valued Logics

Revenge paradoxes are arguments to the effect that a solution to the semantic paradoxes itself generates new paradoxes. Let T be a non-trivial theory of truth. *Qua* non-trivial theory of truth, T avoids the semantic paradoxes, by restricting truth-theoretical or logical principles. A revenge argument against T aims at showing that there are semantic notions that, if formulated in the language of T and characterized by intuitive principles (akin to naïveté for truth), give rise to triviality results similar to the 'standard' paradoxes such as the Liar. Crucially, revenge paradoxes are often justified by the very theories they are directed against: the revenge-breeding semantic notions are usually motivated by some features of T itself, typically definable in T's (classical) meta-theory.

For concreteness, let's focus on a specific revenge-breeding notion: *bivalent determinateness*. Consider the Liar sentence λ. In no Kripke-Kleene valuation is λ assigned value 1. There is, therefore, a (classical) sense in which λ is *not* true. However, this claim cannot be expressed in the object-language, for in every Kripke-Kleene valuation the sentence $\neg\mathsf{Tr}(\ulcorner\lambda\urcorner)$—which formalizes this claim—also receives value 1/2. Attempts to capture the status of λ (and relevantly similar sentences) fail because no connective is definable in strong Kleene semantics that delivers the desired value. However, it is possible to draw this distinction—between sentences that have value 1 and sentences that don't—in our (classical) meta-theory: we have just informally done so. Therefore, a suitable operator can be added to the language to express the desired distinction, and the untrue status of λ. Let D (for 'determinately') be the unary operator governed by the truth-table in Fig. 1. If φ is assigned value 0 or 1/2 in a strong Kleene valuation, $\neg\mathsf{D}(\varphi)$ has value 1, thus formally expressing the idea that φ is not determinate, or not classically, fully, strictly (what have you) true.

However, a Liar-like paradox immediately arises if we combine determinateness with naïve truth. It is easy to see that combining strong Kleene negation with D yields,

φ	$\mathsf{D}(\varphi)$
1	1
$1/2$	0
0	0

Fig. 1 Truth table for the determinateness operator

essentially, classical negation. Therefore, a Liar sentence formulated involving $\neg\mathsf{D}$ (rather than simply $\neg$) is incompatible with naïve truth. Here is a semantic presentation of the paradox.[15]

Example 2.5 (The Determinateness Liar Paradox) Suppose there is a Kripke-Kleene valuation v s.t. it interprets D as in Fig. 1. Let λ_{D} be equivalent to $\neg\mathsf{D}(\mathsf{Tr}(\lambda_{\mathsf{D}}))$. Since v is a strong Kleene valuation, either $v(\lambda_{\mathsf{D}}) = 1$, or $v(\lambda_{\mathsf{D}}) = 1/2$, or $v(\lambda_{\mathsf{D}}) = 0$.

- If $v(\lambda_{\mathsf{D}}) = 1$, then $v(\mathsf{Tr}(\lambda_{\mathsf{D}})) = 1 = v(\mathsf{D}(\mathsf{Tr}(\lambda_{\mathsf{D}})))$. Hence $v(\neg\mathsf{D}(\mathsf{Tr}(\lambda_{\mathsf{D}}))) = 0 = \lambda_{\mathsf{D}}$, which is impossible.
- If $v(\lambda_{\mathsf{D}}) = 1/2$, $v(\mathsf{Tr}(\lambda_{\mathsf{D}})) = 1/2$, $v(\mathsf{D}(\mathsf{Tr}(\lambda_{\mathsf{D}}))) = 0$ and $v(\neg\mathsf{D}(\mathsf{Tr}(\lambda_{\mathsf{D}}))) = 1 = v(\lambda_{\mathsf{D}})$ follow, which is impossible.
- If $v(\lambda_{\mathsf{D}}) = 0$, then $v(\mathsf{Tr}(\lambda_{\mathsf{D}})) = 0 = v(\mathsf{D}(\mathsf{Tr}(\lambda_{\mathsf{D}})))$. Hence $v(\neg\mathsf{D}(\mathsf{Tr}(\lambda_{\mathsf{D}}))) = 1 = \lambda_{\mathsf{D}}$, which is impossible.

This revenge paradox has an immediate consequence:

Proposition 2.6 *No Kripke model can be expanded to a model that* sst-*,* ttt-*,* tst-*, or* stt-*validates both intersubstitutivity of truth and determinateness.*

3 Vagueness in Three-Valued Logics

3.1 Soritical Paradoxes in Three-Valued Logics

We now consider the applications of strong Kleene semantics and our four logics to vague predicates and to soritical paradoxes. Consider a vague predicate P (such as 'tall'), and a countable set $C = \{c_0, c_1, \ldots\}$. Assume that c_0 is a clear case of P, and that $c_0, c_1, \ldots$ are progressively ordered as far as the application of P goes: c_0 is the clearest case of P, c_1 is the clearest case of P after c_0, and so on. Finally, assume that there is a c_j which is a borderline case of P, and that there is an n such that c_n is a clear case of not-P. We now encode these assumptions in a three-valued model $\mathcal{M} = \langle M, f\rangle$ and the valuation $v_{\mathcal{M}}$ based on it.

(a) $v_{\mathcal{M}}(\mathsf{P}(c_0)) = 1$.
(b) There is an individual c_j s.t. $v_{\mathcal{M}}(\mathsf{P}(c_j)) = 1/2$.
(c) There is an individual c_n s.t. $v_{\mathcal{M}}(\mathsf{P}(c_n)) = 0$.

[15] Some authors have argued that revenge paradoxes (of this kind) are not genuine paradoxes. We find that criticism to be misguided, but we will not enter the debate here, for reasons of space. We work under the assumption that revenge paradoxes are genuine paradoxes. See [23] and [67] for discussion.

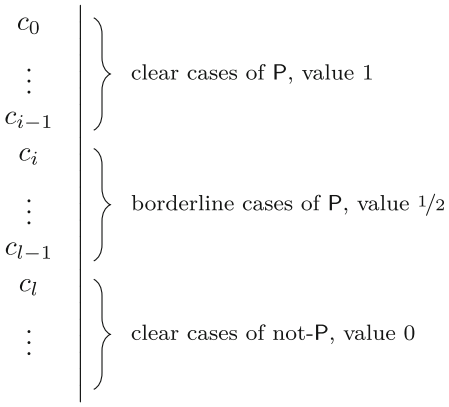

Fig. 2 A soritical model

(d) For every q, $v_{\mathcal{M}}(c_q \sim_{\mathsf{P}} c_{q+1}) = 1$.
(e) $v_{\mathcal{M}}(\mathsf{P}(c_q)) \geq v_{\mathcal{M}}(\mathsf{P}(c_r))$ just in case $q \leq r$.

Call a three-valued model and valuation that respects all of (a)-(e) *soritical*. Application of vague predicates respects always the same pattern in soritical models:

(i) By (c) and the least number principle (LNP), there is a smallest l s.t. $v_{\mathcal{M}}(\mathsf{P}(c_l)) = 0$. By (a), (b), and (e), $0 < j < l \leq n$. By (e), for every $m \geq l$, $v_{\mathcal{M}}(\mathsf{P}(c_m)) = 0$.
(ii) By (b) and LNP, there is a smallest i s.t. $v_{\mathcal{M}}(\mathsf{P}(c_i)) = 1/2$. By (a) and (e), $0 < i \leq j < l$. By (e), for every k s.t. $i \leq k < l$, $v_{\mathcal{M}}(\mathsf{P}(c_k)) = 1/2$.
(iii) By (a) and (e), for any q, if $0 \leq q < i$, then $v_{\mathcal{M}}(\mathsf{P}(c_q)) = 1$.

We can visualize (i)-(ii) as in Fig. 2.

We now use soritical models and valuations to specify theories of vagueness, which employ our four three-valued logics.

Definition 3.1 ssv, ttv, stv, and tsv

- $\Gamma \models_{\mathsf{ssv}} \varphi$ if for every soritical model $\mathcal{M}$ and every induced valuation $v_{\mathcal{M}}$, if $v_{\mathcal{M}}$ makes all the sentences in Γ s-true, it also makes φ s-true.
- $\Gamma \models_{\mathsf{ttv}} \varphi$ if for every soritical model $\mathcal{M}$ and every induced valuation $v_{\mathcal{M}}$, if $v_{\mathcal{M}}$ makes all the sentences in Γ t-true, it also makes φ t-true.
- $\Gamma \models_{\mathsf{tsv}} \varphi$ if for every soritical model $\mathcal{M}$ and every induced valuation $v_{\mathcal{M}}$, if $v_{\mathcal{M}}$ makes all the sentences in Γ t-true, it also makes φ s-true.
- $\Gamma \models_{\mathsf{stv}} \varphi$ if for every soritical model $\mathcal{M}$ and every induced valuation $v_{\mathcal{M}}$, if $v_{\mathcal{M}}$ makes all the sentences in Γ s-true, it also makes φ t-true.

ssv, ttv, tsv, and stv share the same soritical models, but their logical differences induce differences in the principles they satisfy about vagueness, as the next result shows.

Proposition 3.2

– ttv *and* stv *are tolerant logics. For every vague predicate* P*:*

$$\models_{\mathsf{ttv}} \forall x \forall y(\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y)) \qquad \models_{\mathsf{stv}} \forall x \forall y(\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y))$$

– ssv *and* tsv *are intolerant logics. For every vague predicate* P*:*

$$\not\models_{\mathsf{ssv}} \forall x \forall y(\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y)) \qquad \not\models_{\mathsf{tsv}} \forall x \forall y(\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y))$$

Notice that nothing changes if (TOLERANCE- INF) or (TOLERANCE- META) are considered instead: ssv and tsv remain intolerant logics. This concludes our presentation of three-valued logics applied to vague predicates: it is easy to apply the results in this section to show how each of ssv, ttv, tsv, and stv blocks the paradoxical reasoning introduced in 1.3.

3.2 Higher-Order Vagueness in Three-Valued Logics

Higher-order vagueness is also characterized as the paradoxical form of vagueness that affects a given (non-classical) theory of vagueness, and the corresponding solution to soritical paradoxes. Consider a theory of vagueness V. *Qua*, non-trivial theory of vagueness, V avoids the soritical paradoxes, by restricting tolerance or classical logic (or both). A HOV argument against V aims at showing that there are sentences in the meta-theory of V that are affected by the same vagueness that is displayed by object-linguistic sentences. In other words, the meta-theory of V also gives rise to soritical paradoxes, much like ordinary vague expressions do in the object-language.

The similarities between revenge and HOV paradoxes are striking. Both seem to be generated by the same dynamics: a mismatch between a non-classical object-theory and a classical meta-theory. In the case of revenge, the mismatch reinstates essentially the same form of paradoxicality that shows classical logic to be incompatible with naïveté. How about higher-order vagueness? Consider a borderline case of a vague predicate P. We have seen that, using strong Kleene valuations, we cannot characterize the status of a sentence with value $1/2$. Let's then add again bivalent determinateness to our language. A HOV-sorites immediately arises:

1. $\mathsf{D}(\mathsf{P}(c_0))$	[Premiss 1]
2. $c_0 \sim_{\mathsf{D}(\mathsf{P})} c_1$	[Premiss 2]
3. $\mathsf{D}(\mathsf{P}(c_0)) \wedge c_0 \sim_{\mathsf{D}(\mathsf{P})} c_1$	[1, 2, $\wedge$-I]
4. $\forall x \forall y(\mathsf{D}(\mathsf{P}(x)) \wedge x \sim_{\mathsf{D}(\mathsf{P})} y \rightarrow \mathsf{D}(\mathsf{P}(y))$	[TOLERANCE]
5. $\mathsf{D}(\mathsf{P}(c_0)) \wedge c_0 \sim_{\mathsf{D}(\mathsf{P})} c_1 \rightarrow \mathsf{D}(\mathsf{P}(c_1))$	[3, $\forall$-E]
6. $\mathsf{D}(\mathsf{P}(c_1))$	[3, 5, $\rightarrow$-E]
7. $\vdots$	[reiterate the above passages, starting with 6]
$\mathsf{D}(\mathsf{P}(c_n))$	

A model-theoretical presentation can be easily extracted from this description. Since only tt and st are tolerant (Proposition 3.2), it is instructive to observe how

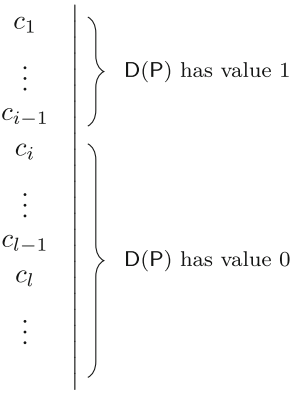

Fig. 3 A soritical model for determinateness

they would treat the above paradox. Let $\mathcal{M}$ be a soritical model. By construction, $v_{\mathcal{M}}(\mathsf{D}(\mathsf{P}(c_0))) = 1 = v_{\mathcal{M}}(\mathsf{D}(\mathsf{P}(c_1))) = \ldots = v_{\mathcal{M}}(\mathsf{D}(\mathsf{P}(c_{i-1})))$. Moreover, for every c_j s.t. $j \geq i$, $v_{\mathcal{M}}(\mathsf{D}(\mathsf{P}(c_j))) = 0$. The model can be visualized in Fig. 3.

Clearly, $\mathsf{D}(\mathsf{P}(c_{i-1})) \wedge c_{i-1} \sim_{\mathsf{D}(\mathsf{P})} c_i \rightarrow \mathsf{D}(\mathsf{P}(c_i))$ has value 0, and therefore TOLERANCE is not generally tolerantly valid. More generally:

Proposition 3.3 *No soritical model can be expanded to a model that* ssv-*,* ttv-*,* tsv-*, or* stv-*validates both tolerance and determinateness.*

One could object to the legitimacy of the above HOV-sorites as follows: the claim that any two c_m and c_{m+1} are similar with respect to $\mathsf{D}(\mathsf{P})$ is not justified, because D precisely distinguishes between cases of P that, while similar, fall into different semantic categories. Therefore, $c_m \sim_{\mathsf{D}(\mathsf{P})} c_{m+1}$ does not generally hold, and the corresponding case of TOLERANCE isn't justified either. Here's our reply. The objection is based on the semantics of D, an operator intended to capture 'determinately'. Even if we make a technical use of 'determinately' to distinguish clear from borderline cases of 'tall' in a model, 'determinately tall' remains a vague predicate of natural language, and a tolerant one too, as anyone's linguistic competence can attest. Therefore, we are not entitled to reject tolerance because of its conflict with the intended, pre-theoretic interpretation of a predicate: otherwise, we could reject tolerance for 'tall' on the basis that it is incompatible with the classical semantics for 'tall'. So, either we reject tolerance *tout court*, or we accept it for all the predicates which are intuitively tolerant in natural languages. What is unjustified is accepting tolerance only for *some* vague predicates, just because accepting it across the board conflicts with our theory. Rather, that our theory cannot accommodate it indicates that it hasn't blocked soritical paradoxes in all their breadth: soritical paradoxes resurface as soon as some semantic notions, such as determinateness, are made explicit – a typical revenge dynamics.

Before moving on, note that we have presented Kripkean fixed-point models and soritical models separately, focusing first on the truth-theoretical fragment of $\mathcal{L}_{\mathrm{t,v}}$ and then on its vague fragment. We did so to keep the presentation simple, but the two models can be combined with no effort. One simply starts with a soritical model—that gives to atomic sentences of the form $\mathsf{P}(c)$ value 1, 1/2, or 0—and builds a Kripke

fixed point over it. However, giving an explicitly unified interpretation to both the truth-theoretical and the vague parts of $\mathcal{L}_{t,v}$ was not important for presenting Kripke and soritical models. It will become important just now.

4 Unifying the Paradoxes

Although paradoxes of truth and vagueness can be loosely connected via generic features (they are both arguments to triviality, or non-existence results), until now they remain heterogeneous phenomena: they have different catalysts, and there is no obvious similarity between their fine-grained structure. Yet, some clues suggest a deeper similarity between the paradoxes. For instance, in both the revenge and HOV cases, the catalyst of the paradox—bivalent determinateness—and the reasoning that leads to the paradoxical arguments—expressing classical (meta-theoretical) notions in a non-classical (object-)theory—are the same. This suggests that revenge and HOV paradoxes obtain by applying the same pattern of reasoning 'on top' of the corresponding paradoxes. Should the latter be identified (in some sense to be made precise), such identification would carry over to the revenge and HOV case.

In this part of the paper, we substantiate the idea that paradoxes of truth and vagueness are deeply related, by pursuing a two-fold strategy. First, in this section, we formalize and analyze the paradoxes in a suitable unified setting. This requires extending the traditional notion of model in such a way that paradoxical reasoning, rather than showing that certain models do not exist, can indeed be modeled. We do so by generalizing an *equational semantics*, originally developed in [68], where sentences are assigned equation systems rather than simple numerical values. We then develop our equational semantics into a full-fledged notion of equational consequence, tailored to analyze arguments that lead to contradiction via uses of naïveté and tolerance. The upshot is that paradoxical arguments, including revenge and HOV-paradoxes, are shown to display the same kind of (problematic) reasoning pattern, which becomes visible to an equational analysis. For this reason, we call this a *unification* of the paradoxes of truth and vagueness.

In Section 5, we push our unification further: we single out a principle of *indiscernibility*, and show that naïveté and tolerance are special cases of it. We call this a *reduction* of the paradoxes of truth and vagueness. Unification and reduction provide us with two different but related answers to the question whether semantic and soritical paradoxes should be identified. On the basis of what we call unification, semantic and soritical paradoxes can be identified in the sense they can be thought of as 'similar' arguments—our model provides the relevant sense of similarity. On the basis of what we call reduction, semantic and soritical paradoxes can be identified in the sense that naïveté and tolerance both derive from the same principle, i.e. indiscernibility.

4.1 Heuristics

Let us now explain the basic idea of our equational semantics, mostly via examples. Consider an arbitrary $\mathcal{L}_{t,v}$-sentence φ. If φ is an atomic, non-semantic sentence (i.e., an atomic sentence which is not a truth-predication), then its semantic value is determined

by the base model $\mathcal{M}$ we are considering. If φ is a complex sentence, its value depends on the value of its sub-formulae, as per Definition 1.5:

- if φ is $\neg\psi$, the value of it is $1-$ the value of ψ,
- if φ is $\psi \wedge \chi$, the value of it is the minimum of the values of ψ and χ,
- if φ is $\forall x \psi(x)$, the value of it is the infimum of the values of its instances $\psi(t)$.

The above clauses can be used to define valuation functions – as in Definition 1.5 – but also *equation systems*. That is, we can write them as

- $v = 1 - v_1$
- $v = \min(v_1, v_2)$
- $v = \inf(w_1, w_2, \ldots)$

for v the value of φ, v_1 the value of ψ, v_2 the value of χ, and $w_1, w_2, \ldots$ the values of $\psi(t_1), \psi(t_2), \ldots$ Since strong Kleene semantics is compositional, this process goes on: when we have associated equations with φ, we associate equations to its sub-formulae $\psi_1, \ldots, \psi_n, \ldots$, and then we associate equations with each of the latter, and so on.

Let's consider one example. Suppose that $\mathcal{L}_{\mathrm{t,v}}$ contains some arithmetical vocabulary, interpreted in the usual way, and consider the sentence $0 = 1 \rightarrow (0+n) \neq (1+n)$. By re-writing it using the official connectives of $\mathcal{L}_{\mathrm{t,v}}$, it becomes $\neg[0 = 1 \wedge \neg(0+n) \neq (1+n)]$, which yields the following equation system:

$$\begin{cases} v = 1 - w \\ w = \min(v_1, v_2) \\ v_1 = 0 \\ v_2 = 1 - v_3 \\ v_3 = 1 \end{cases}$$

where v_3 stands for the value of $(0+n) \neq (1+n)$, v_2 for the value of $\neg(0+n) \neq (1+n)$, v_1 for the value of $0 = 1$, w the value of $0 = 1 \wedge \neg(0+n) \neq (1+n)$, and v for the value of $\neg[0 = 1 \wedge \neg(0+n) \neq (1+n)]$. The equations $v_1 = 0$ and $v_3 = 1$ are justified by the assumption that the arithmetical vocabulary $(0, 1, +)$ is interpreted in the usual way. The system has a unique solution, namely $v = v_3 = 1$ and $w = v_1 = v_2 = 0$, as expected.

As the example shows, the target sentence φ is iteratively decomposed into its subsentences until undecomposable elements are reached. The resulting sentences determine an equation system, which follows the strong Kleene clauses. Finally, the possible solutions of the systems are considered, to provide a value to all the subsentences of φ and to φ itself. The solutions depend on the logical form of φ and on the base model.

Until here, we just re-wrote the defining equations of a strong Kleene valuation induced by a base model. We neglected the truth predicate, the vague vocabulary, and the determinateness operator. Let's start with truth. As above, we re-write the semantics for truth-predications, provided by Kripke-Kleene valuations (Definition 2.1, Lemma 2.2) in equational terms:

- the value of $\mathsf{Tr}(\ulcorner \psi \urcorner)$ is the value of ψ.

The analysis of simple cases, such as $\mathsf{Tr}(\ulcorner 0 = 1 \rightarrow (0 + n) \neq (1 + n)\urcorner)$, is now immediate. Let's now look at a more interesting example, such as a Liar sentence λ. λ is the sentence $\neg\mathsf{Tr}(t_\lambda)$, for a term t_λ that (in the selected model) denotes the same element as $\ulcorner\neg\mathsf{Tr}(t_\lambda)\urcorner$. Therefore, we associate the equation $v = 1 - w$ with $\neg\mathsf{Tr}(t_\lambda)$, where v is the variable for the value of $\neg\mathsf{Tr}(t_\lambda)$, and w for the value of $\mathsf{Tr}(t_\lambda)$. Moreover, we associate the equation $w = v$ with $\mathsf{Tr}(t_\lambda)$, in line with the above clause to evaluate truth-predications. We therefore have the following system:

$$\begin{cases} v = 1 - w \\ v = w \end{cases}$$

which has only one solution: $v = 1/2 = w$. A Liar sentence λ is associated with an equation system which directly expresses that, given naïveté, λ must have the same value as its negation, and that its only possible value is $1/2$. Example of similar analyses of paradoxical sentences – such as Curry's sentences of McGee's sentences – readily multiply.

Does every equation system defined in this way have a unique solution? No. Counterexamples, interestingly, correspond to specific kinds of paradoxical sentences. Consider a *truth-teller sentence* τ, which abbreviates $\mathsf{Tr}(\ulcorner t_\tau \urcorner)$, where the term t_τ denotes $\ulcorner\mathsf{Tr}(\ulcorner t_\tau \urcorner)\urcorner$. τ (informally) says of itself that it is true. τ is easily seen to be associated with the equation system consisting only of $v = v$, which clearly has as many solutions as there are semantic values. On the other hand, the sentences employed in *revenge paradoxes* yield equation systems with no solutions. As the reader can easily check, the Determinateness Liar sentence λ_D yields an equation system with no solutions in our value space $\{1, 1/2, 0\}$.

Rossi (2019) shows that every sentence of languages such as our $\mathcal{L}_{\mathrm{t,v}}$ can be associated with an equation system of the above kind (via a suitable model-theoretic construction). Such equation systems (as every equation system) give rise to three possibilities: either it has a unique solution (in the selected value space), or it has more than one solutions, or it has no solutions. Therefore, the solvability of equation system is used, in [68], to provide an exhaustive characterization of *sentences* within one single model:

- sentences whose equation admits a unique classical solution (i.e., 0 or 1) are deemed *classical* (example: $\mathsf{Tr}(\ulcorner 0 = 0\urcorner)$);
- sentences whose equation admits a unique non-classical solution (i.e., $1/2$) are deemed *Liar-like* (example: λ)
- sentences whose equation admits more than one solution are deemed *Truth-teller-like* (example: τ).
- sentences whose equation admits no solution are deemed *revenge-like* (example: λ_D).

Here, we are not concerned with a taxonomy of paradoxical sentences, but with the structure of paradoxical reasoning. Therefore, we will extend the semantics of [68] in order to cover the vague vocabulary as well, and then extend it from sentences to *arguments*.

4.2 Equational Semantics

Let N_3 be our three-valued numerical value space, i.e., $\mathsf{N}_3 = \{0, 1/2, 1\}$. We now need a language to assign equations to formulas of $\mathcal{L}_{\mathrm{t,v}}$.

Definition 4.1 Let $\mathcal{L}_3$ be the language whose alphabet includes:

- a set Var_3 of variables $\{v_{\varphi_1}, \ldots, v_{\varphi_n}, \ldots\}$, where each φ_k is the k-th element in a non-redundant enumeration of sentences of $\mathcal{L}_{\mathrm{t,v}}$,[16]
- a set Con_3 containing an individual constant for every element in N_3,
- a binary relation $=$ for equality.

We now define terms and atomic formulae of $\mathcal{L}_3$, in order to formally define equations and equation systems. Equations are just atomic formulae of the form $s = t$, for s, t $\mathcal{L}_3$-terms, and equation systems are sets of equations.

Definition 4.2 The set of $\mathcal{L}_3$-terms is defined by recursively closing Var_3 and Con_3 under the following operations:[17]

- $(1 - x)$,
- $\min(x, y)$,
- $\inf\{x_1, x_2, \ldots, x_n, \ldots\}$.

Atomic $\mathcal{L}_3$-formulas are expressions of the form $s = t$, for $\mathcal{L}_3$-terms s and t.

Some notation: Let $\mathbb{E}_3$ denote the set of atomic $\mathcal{L}_3$-formulas, $\mathbf{e}$ (possibly with indices) vary over elements of $\mathbb{E}_3$, $\mathbf{E}$ (possibly with indices) vary over elements of $\mathcal{P}(\mathbb{E}_3)$, and let $\mathsf{Var}(\mathbf{E})$ indicate the collection of $\mathcal{L}_3$-variables appearing in $\mathbf{E}$. We write $\mathbf{E}_\varphi$ for the equation system defined by φ (as shown in [68], every sentence φ is associated exactly with one equation system). Finally, the $\mathcal{L}_3$-variable assigned to φ in $\mathbf{E}_\varphi$ is called the *principal variable* of $\mathsf{Var}(\mathbf{E}_\varphi)$, and will be denoted with v_φ.

The elements of $\mathbb{E}_3$ are the equations definable from the strong Kleene valuation clauses from Definition 1.5. As in [68], these equations are assigned to $\mathcal{L}_{\mathrm{t,v}}$-formulae, in a way that mimics the strong Kleene schema. The following definition provides a semantics for $\mathcal{L}_{\mathrm{t,v}}$-sentences, which gives them both numerical values (as usual) *and* equations.

Definition 4.3 A *equational structure* for $\mathcal{L}_{\mathrm{t,v}}$ is a structure S_3 given by

$$S_3 = \langle \mathcal{M}, \mathsf{Con}_3, \mathbb{E}_3, e, A \rangle$$

with $\mathcal{M}$ a soritical, acceptable $\mathcal{L}_{\mathrm{t,v}}$-structure, Con_3 and $\mathbb{E}_3$ as above, and s.t.:

- e is a *valuation function* $e : \mathsf{Sent}_{\mathcal{L}_{\mathrm{t,v}}} \longmapsto \mathcal{P}(\mathbb{E}_3)$ from $\mathcal{L}_{\mathrm{t,v}}$-sentences to equations, obeying the clauses of Definition 1.5 and naïveté for truth;
- A is a set of partial functions $\alpha : \{\mathsf{Var}(\{\mathbf{e}\}) \mid \mathbf{e} \in \mathbb{E}_3\} \longmapsto \mathsf{Con}_3$. That is, each α is an assignment of values in Con_3 to variables in $\mathsf{Var}(\{\mathbf{e}\})$.

[16] We often omit the subscript φ_i for simplicity.

[17] Clearly, they match the operation employed in Definition 1.5.

We omit the model-theoretic construction of the function e and of the set A (we refer the reader to [68, Sections 3-4] for details). Informally, e works as in the presentation given at the beginning of Section 4: e associates with an $\mathcal{L}_{\mathrm{t,v}}$-formula φ its equation system $\mathbf{E}_\varphi$. More precisely:

- if φ is not a non-semantic atomic formula, then $\mathbf{E}_\varphi$ includes the equation describing how the semantic value of φ in N_3 is calculated, depending on the logical form of φ;
- $\mathbf{E}_\varphi$ also contains the equations associated with the values of the subformulae of φ, and for their subformulae, and so on until non-semantic atomic formulae are reached; the latter are assigned equations of the form $v = 0$ or $v = 1$, depending on whether they receive value 0 or 1 in the base model $\mathcal{M}$.[18]

To summarize: with every formula φ of $\mathcal{L}_{\mathrm{t,v}}$ we associate a system of equations $e(\varphi)$ in $\mathbb{E}_3$, which contains the fundamental semantic 'information' required to assign to φ its truth value in S_3. It is then possible to check whether $\mathbf{E}_\varphi$ admits solutions in N_3 or not.

The following result from [68] guarantees that Definition 4.3 works as in the informal examples at the beginning of the section.

Proposition 4.4 *For every acceptable $\mathcal{L}_{\mathrm{t,v}}$-structure $\mathcal{M}$:*

(i) There exists a non-empty set of equational structures,
(ii) If $S_3' = \langle \mathcal{M}, \mathsf{Con}_3, \mathbb{E}_3, e', A' \rangle$ and $S_3'' = \langle \mathcal{M}, \mathsf{Con}_3, \mathbb{E}_3, e'', A'' \rangle$ are two equational structures generated by $\mathcal{M}$, then $e' = e''$.
(iii) The set of equational structures generated by $\mathcal{M}$ has a $\subseteq$-least element.

Item (i) guarantees that Definition 4.3 is not vacuous: there are equational structures. Item (ii) states that the function which assigns $\mathcal{L}_3$-equations to $\mathcal{L}_{\mathrm{t,v}}$-sentences is unique. Finally, item (iii) together with item (ii) shows that the set A is the only set that can possibly differentiate two equational structures generated by $\mathcal{M}$ and, therefore, there is a least set of solutions to the $\mathcal{L}_3$-equations assigned to $\mathcal{L}_{\mathrm{t,v}}$-sentences.[19]

Equational structures endow 'traditional' models with sets of equations and their solutions. What is the advantage in interpreting $\mathcal{L}_{\mathrm{t,v}}$-sentences via equational structures? The answer becomes clear if we consider the model-theoretic presentation of semantic (Section 1.2), soritical (Section 1.3), revenge (Section 2.2), and HOV paradoxes (Section 3.2). Model-theoretically viewed, paradoxes are *non-existence results*: they show that certain valuations (e.g., classical valuations obeying naïveté) do not exist. Therefore, the relevant semantic notions—naïveté, tolerance, or bivalent determinateness—cannot simply be modeled in traditional models. As a consequence, arguments involving these notions (regardless of whether they are paradoxical or not) cannot be formally represented there. On the other hand, equational structures that interpret naïve truth, tolerance, or bivalent determinateness are unproblematically

[18] Some assignments of equations never 'reach' atomic formulae—in the Visser-Yablo paradox, for example, we have an infinitely descending chain of truth-predications. However, the construction of e does not require that the structure of the components of a given sentences is well-founded.

[19] As per the construction in [68], such set is given by only solving systems that admit exactly one solution, and not solving systems that admit more than one solutions.

available (Proposition 4.4), and easily accommodate paradoxical sentences. Paradoxical sentences are interpreted in an equational structure as equation systems which do not have a unique classical solution. Having a suitable interpretation of sentences in equational structures, the next step is to extend it to *arguments*. This is the topic of the next section.

4.3 Equational Consequence

Here, we employ equational structures to define equational notions of *consequence*, and model arguments involving naïve truth, tolerance, or bivalent determinateness. Let us fix some more notation. For every assignment α and for every $\mathbf{e}$ in $\mathbb{E}_3$, let $\models^{\alpha} \mathbf{e}$ indicate that $\mathbf{e}$ is a true arithmetical equation under the assignment α of values in N_3 to $\mathsf{Var}(\{\mathbf{e}\})$. So, $\models^{\alpha} \mathbf{e}$ holds if $\alpha(\mathsf{Var}(\{\mathbf{e}\}))$ is a true arithmetical identity. Let us also put, for every assignment α and for every $\mathbf{E} \subseteq \mathbb{E}_3$, $\alpha(\mathbf{E}) = \{\alpha(\mathsf{Var}(\{\mathbf{e}\})) \mid \mathbf{e} \in \mathbf{E}\}$, and put $\models^{\alpha} \mathbf{E}$ if and only if $\models^{\alpha} \mathbf{e}$ for every $\mathbf{e} \in \mathbf{E}$. We can now use the existence of solutions to $\mathcal{L}_3$-equations to provide a generalized notion of satisfiability, which we will use to model paradoxical arguments:

Definition 4.5 Let $S_3 = \langle \mathcal{M}, \mathsf{Con}_3, \mathbb{E}_3, e, A \rangle$ be an equational structure.

- A set $\mathbf{E} \subseteq \mathbb{E}_3$ is *solvable in* S_3 if and only if there exists an assignment $\alpha \in A$ such that $\models^{\alpha} \mathbf{E}$.
- An $\mathcal{L}_{\mathrm{t,v}}$-sentence φ is *satisfiable in* S_3 if and only if $\mathbf{E}_{\varphi}$ is solvable.
- A set Γ of $\mathcal{L}_{\mathrm{t,v}}$-sentences is *satisfiable in* S_3 if and only if there is an assignment $\alpha \in A$ such that $\models^{\alpha} \mathbf{E}_{\varphi}$, for every $\varphi \in \Gamma$ (i.e. if every formula of Γ is satisfied by one and the same assignment).

This notion of satisfiability can be specified in order to recover the notion of strict and tolerant truth, introduced in Section 1.4.

Definition 4.6 Let $S_3 = \langle \mathcal{M}, \mathsf{Con}_3, \mathbb{E}_3, e, A \rangle$ be an equational structure.

- An $\mathcal{L}_{\mathrm{t,v}}$-sentence φ is *strictly (tolerantly) true in* S_3 ($\mathsf{s}(\mathsf{t})$-true), if there is an $\alpha \in A$ s.t. $\alpha(v_{\varphi}) = 1$ ($\alpha(v_{\varphi}) \geq 1/2$) and $\models^{\alpha} \mathbf{E}_{\varphi}$.
- A set Γ of $\mathcal{L}_{\mathrm{t,v}}$-sentences is *strictly (tolerantly) true in* S_3 ($\mathsf{s}(\mathsf{t})$-true), if there is an $\alpha \in A$ s.t. $\alpha(v_{\varphi}) = 1$ ($\alpha(v_{\varphi}) \geq 1/2$) and $\models^{\alpha} \mathbf{E}_{\varphi}$, for every $\varphi \in \Gamma$.

We can now use the above definition to specify notions of mne-*satisfiability in* S_3 for arguments, where $\mathsf{m}, \mathsf{n} \in \{\mathsf{s}, \mathsf{t}\}$, as follows.

Definition 4.7 (Equational validity/consequence) Let $\Gamma \cup \{\varphi\}$ be a set of $\mathcal{L}_{\mathrm{t,v}}$-sentences.

- $\Gamma \models_{\mathsf{sse}} \varphi$ if, for every equational structure S_3, every assignment α in S_3 that makes Γ s-true, makes also φ s-true;
- $\Gamma \models_{\mathsf{tte}} \varphi$ if, for every equational structure S_3, every assignment α in S_3 that makes Γ t-true, makes also φ t-true;
- $\Gamma \models_{\mathsf{tse}} \varphi$ if, for every equational structure S_3, every assignment α in S_3 that makes Γ t-true, makes also φ s-true;

– $\Gamma \models_{\mathsf{ste}} \varphi$ if, for every equational structure S_3, every assignment α in S_3 that makes Γ s-true, makes also φ t-true.

sse, tte, tse, and ste are patterned after ss, tt, ts, and st (Definition 1.7), but with a few differences. First, the semantic values (1, 0, and 1/2) employed to determine whether φ equationally mne-follows from Γ are results of equations. As such, the possibility of equations *not* admitting solutions is explicitly incorporated into the notion of consequence, thereby making it possible to reproduce paradoxical reasonings. Second, equational structures, by design, are defined over soritical models and incorporate naïveté for the valuation of truth-predications. Therefore, vague atomic sentences and truth-predications are not treated as arbitrary atomic sentences, and arbitrarily interpreted by any given semantic structure. So, sse, tte, tse, and ste are perhaps best seen as incorporating the notions of consequence of Definitions 2.3 and 3.1, combining them in an equational framework where paradoxical arguments that cannot be modelled in the latter (such as revenge and HOV arguments) can be fully represented.

We now put sse, tte, tse, and ste at work, and see how they provide a unifying analysis of semantic and soritical paradoxes, and of revenge and HOV paradoxes.

4.4 Semantic and Soritical Paradoxes as Equational Arguments

In this section, we show that our theories of equational consequences $\models_{\mathsf{mne}}$ reproduce the treatment of paradoxes in the corresponding theories of truth (mnt) and of vagueness (mnv). We start from the Liar case. First, notice that all of the main ingredients of the Liar reasoning, i.e. the naïveté of truth and diagonalization (in the form of $\text{MD}_{\text{IAG}_\lambda}$), have been incorporated in the notion of equational structure. Let then Γ_λ be $\{\lambda\}$.

Proposition 4.8 $\Gamma_\lambda \models_{\mathsf{mne}} \bot$ *only vacuously if* m=s, *and* $\Gamma_\lambda \not\models_{\mathsf{mne}} \bot$ *if* m=t.

Proof Let S_3 be an equational structure. Definition 4.7 requires that there is an assignment α that makes Γ_λ m-true. If m=s, this would require α to assign value 1 to v_λ. However, this cannot be the case, since $\mathbf{E}_\lambda = \{v_\lambda = 1 - w, w = v_\lambda\}$, which admits only value 1/2 as solution. Therefore, there is no such α, which means that $\Gamma_\lambda \models_{\mathsf{sne}} \bot$ holds vacuosly for every $\mathsf{n} \in \{\mathsf{s}, \mathsf{t}\}$.

Let m=t and let α be an assignment that makes Γ_λ t-true (which happens whenever $\alpha(v_\lambda) = 1/2$). For the argument to be tne-valid, it is required that $\bot$ is n-true in α, i.e. that $\alpha(v_\bot) = 1$ if $\mathsf{n} = \mathsf{s}$, and that $\alpha(v_\bot) \geq 1/2$ if $\mathsf{n} = \mathsf{t}$. However, $\alpha(v_\bot) = 0$, and therefore neither of the two conditions can be met. □

Let us now turn to the Sorites Paradox. The argument involves a vague predicate P of $\mathcal{L}_{\mathrm{t,v}}$, a clear-cut case in which P holds, a_0, and a clear-cut case in which P does not hold, a_n. Then, a contradiction arises by suitably applying all the instances of the tolerance principle involving the relation of P-similarity $\sim_{\mathsf{P}}$. We can formalize the argument in our framework as follows. Let Γ_σ be the following set of sentences of

$\mathcal{L}_{\mathrm{t,v}}$:

$$\Gamma_\sigma = \left\{ \begin{array}{l} \mathsf{P}(a_0), \\ \mathsf{P}(a_0) \wedge a_0 \sim_{\mathsf{P}} a_1 \rightarrow \mathsf{P}(a_1), \\ \quad \vdots \\ \mathsf{P}(a_{n-2}) \wedge a_{n-2} \sim_{\mathsf{P}} a_{n-1} \rightarrow \mathsf{P}(a_{n-1}), \\ a_0 \sim_{\mathsf{P}} a_1, \\ \quad \vdots \\ a_{n-1} \sim_{\mathsf{P}} a_n \end{array} \right\}$$

Just as Γ_λ, also Γ_σ encodes the relevant assumptions at play in a soritical argument, as displayed in Example 1.3. We can now prove the following:

Proposition 4.9 $\Gamma_\sigma \models_{\mathsf{mne}} \mathsf{P}(a_n)$ *only vacuously if* $\mathsf{m{=}s}$, *and* $\Gamma_\sigma \not\models_{\mathsf{mne}} \mathsf{P}(a_n)$ *if* $\mathsf{m{=}t}$.

Proof As for Proposition 4.8, we have to show whether the clauses of Definition 4.7 hold or fail. Let then S_3 be an equational structure. For the argument from Γ_σ to $\mathsf{P}(a_n)$ be mne-valid in S_3, it is required that all sentences in Γ_σ be made m-true by an assignment α. If $\mathsf{m{=}s}$ there must be an assignment α in S_3 that makes all the system of equations associated to any sentence in Γ_σ solvable by assigning the value 1 to the variables corresponding to these sentences. However, this is not possible with Γ_σ, since this requires that value 1 is assigned to the variables corresponding to every $\mathsf{P}(a_i)$ to make sentences of the form

$$\mathsf{P}(a_{i-1}) \wedge a_{i-1} \sim_{\mathsf{P}} a_i \rightarrow \mathsf{P}(a_i)$$

s-true. Due to the fact that (i) $a_1, \ldots, a_n$ form a soritical series, and (ii) the model $\mathcal{M}$ of S_3 is soritical, this cannot happen. Therefore there is no such assignment in any equational structure. Hence, the argument from Γ_σ to $\bot$ is vacuously sne-valid for every $\mathsf{n} \in \{\mathsf{s}, \mathsf{t}\}$.

If $\mathsf{m{=}t}$, then there can be an assignment α that makes Γ_σ m-true. Now, since formulas $\mathsf{P}(a_i)$, for $0 \leq i \leq n$, are all atomic, their valuation in S_3 depends on the model $\mathcal{M}$ of S_3. Since $\mathcal{M}$ is soritical, and a_n is a clear-cut case of failure of P, then we have $\alpha(v_{\mathsf{P}(a_n)}) = 0$ which entails that $\mathsf{P}(a_n)$ is not n-true for either $\mathsf{n} = \mathsf{s}$, or $\mathsf{n} = \mathsf{t}$. Hence, the argument is tne-invalid for every $\mathsf{n} \in \{\mathsf{s}, \mathsf{t}\}$. □

4.5 Revenge and HOV Paradoxes as Equational Arguments

We now turn to revenge and HOV paradoxes. Let us start from the Determinateness Liar. The argument involves a determinateness operator D, which is expected to behave as follows: $\mathsf{D}(\varphi)$ has value 1 if φ has value 1, and value 0 otherwise. Consider the language $\mathcal{L}_{\mathrm{t,v,d}} := \mathcal{L}_{\mathrm{t,v}} \cup \{\mathsf{D}\}$. The Determinateness Liar Paradox involves a sentence λ_{D} which is equivalent to $\neg\mathsf{D}(\mathsf{Tr}(\ulcorner\lambda_{\mathsf{D}}\urcorner))$ – where, again, 'equivalent' means that the diagonalization rule ($\mathrm{MDIAG}_{\lambda_{\mathsf{D}}}$) holds for it. Given the semantics of D and the functioning of the valuation e of S_3 (Section 4.2), the sentence λ_{D} yields the following

equation system:

$$\mathbf{E}_{\lambda_{\mathsf{D}}} = \{v = 1 - w, w = 1 - \min(1, 2(1 - v))\},$$

where v and w are variables assigned to λ_{D} and $\mathsf{D}(\mathsf{Tr}(\ulcorner\lambda_{\mathsf{D}}\urcorner))$ respectively. Let $\Gamma^{\mathsf{D}}_{\lambda}$ be $\{\lambda_{\mathsf{D}}\}$.

Proposition 4.10 *For every* $\mathsf{m}, \mathsf{n} \in \{\mathsf{s}, \mathsf{t}\}$, $\Gamma^{\mathsf{D}}_{\lambda} \models_{\mathsf{mne}} \bot$.

Proof Let S_3 be an equational structure, and let $\alpha \in A$. A quick calculation shows that $\mathbf{E}_{\lambda_{\mathsf{D}}}$ is not solvable in N_3.[20] Therefore, for no assignment in S_3, $\Gamma^{\mathsf{D}}_{\lambda}$ can be m-true. Since there is no α making $\Gamma^{\mathsf{D}}_{\lambda}$ m-true, it follows that all of them make $\bot$ n-true, i.e. $\Gamma^{\mathsf{D}}_{\lambda} \models_{\mathsf{mne}} \bot$. □

The HOV case is easily seen to be similar. As in Section 4.4, let P be a vague predicate. We can build a HOV argument along the lines of Section 3.2 applied to $\mathsf{D}(\mathsf{P}(a_0))$, where a_0 is a clear-cut case of P, and suitable instances of $x \sim_{\mathsf{D}(\mathsf{P})} y$, to conclude $\mathsf{D}(\mathsf{P}(a_n))$, where a_n is a clear-cut case of not-P. Here is a formalization of the premises of the argument in $\mathcal{L}_{\mathrm{t,v,d}}$:

$$\Gamma^{\mathsf{D}}_{\sigma} = \left\{ \begin{array}{l} \mathsf{D}(\mathsf{P}(a_0)), \\ \mathsf{D}(\mathsf{P}(a_0)) \wedge a_0 \sim_{\mathsf{D}(\mathsf{P})} a_1 \rightarrow \mathsf{D}(\mathsf{P}(a_1)), \\ \quad \vdots \\ \mathsf{D}(\mathsf{P}(a_{n-1})) \wedge a_{n-1} \sim_{\mathsf{D}(\mathsf{P})} a_n) \rightarrow \mathsf{D}(\mathsf{P}(a_n)), \\ a_0 \sim_{\mathsf{D}(\mathsf{P})} a_1 \\ \quad \vdots \\ a_{n-1} \sim_{\mathsf{D}(\mathsf{P})} a_n \end{array} \right\}$$

As in the previous cases (Γ_{λ}, Γ_{σ}, and $\Gamma^{\mathsf{D}}_{\lambda}$), also $\Gamma^{\mathsf{D}}_{\sigma}$ models the premises of a HOV soritical paradox, as informally presented in Example 3.2.

Proposition 4.11 *For every* $\mathsf{m}, \mathsf{n} \in \{\mathsf{s}, \mathsf{t}\}$, $\Gamma^{\mathsf{D}}_{\sigma} \models_{\mathsf{mne}} \mathsf{D}(\mathsf{P}(a_n))$.

Proof Let then S_3 be an equational structure, and let α be any assignment from the set A of S_3. Since the model $\mathcal{M}$ of S_3 is soritical, there is an i s.t. $0 < i < n$ and s.t. $\alpha(v_{\mathsf{P}(a_j)}) = 1/2$ for every $j \geq i$ with $j < n$. It follows that $\alpha(v_{\mathsf{D}(\mathsf{P}(a_j))}) = 0$, and that the formula

$$\mathsf{D}(\mathsf{P}(a_{i-1})) \wedge a_{i-1} \sim_{\mathsf{D}(\mathsf{P})} a_i \rightarrow \mathsf{D}(\mathsf{P}(a_i))$$

is neither t- nor s-true (recall that in a soritical model sentences of the form $a_{i-1} \sim_{\mathsf{D}(\mathsf{P})} a_i$ have always value 1). Since α was arbitrary, this shows that there is no α that makes $\Gamma^{\mathsf{D}}_{\sigma}$ m-true, and hence that every such α makes $\mathsf{D}(\mathsf{P}(a_n))$ n-true for both $\mathsf{n} = \mathsf{s}$, and $\mathsf{n} = \mathsf{t}$. □

[20] $\mathbf{E}_{\lambda_{\mathsf{D}}}$ is solvable in an extended value space, letting $v = 2/3$ and $w = 1/3$.

	Semantic	Soritical	Revenge	HOV
sse	✓	✓	✗	✗
tte	✓	✓	✗	✗
ste	✓	✓	✗	✗
tse	✓	✓	✗	✗

Fig. 4 Equational theories and paradoxes

Propositions 4.10 and 4.11 illustrate the 'value added' by equational semantics to the usual notion of model: not only it makes it possible to reproduce the three-valued interpretation of the Liar and soritical paradoxes (Section 4.4), but it also provides us with models for the language of revenge and HOV paradoxes, while no extension of a 'traditional' model so to interpret bivalent determinateness is possible in the first place (Propositions 2.6 and 3.3).[21] Our results correctly predict the vulnerability of our four target logics to revenge and HOV paradoxes.

We summarize the results of Sections 4.4-4.5 in the table reproduced in Fig. 4. The first row indicates the formalized paradox in question, while the first column indicates our unified equational theories where the paradoxes are analyzed. The symbol ✓ indicates that the corresponding theory successfully blocks the corresponding paradox, because the formalized paradoxical argument to $\bot$ is either not equationally valid or only vacuously equationally valid in the target theory, while ✗ indicates that the corresponding theory fails to block the corresponding paradox, because the formalized paradoxical argument to $\bot$ is equationally valid in the target theory.

5 Reducing Naïveté and Tolerance to Indiscernibility

As anticipated in the Introduction, we have now obtained a *unification* of semantic and soritical paradoxes: via the equational semantics introduced in Section 4, both kinds of paradoxes (including revenge and HOV paradoxes) have been shown to work in the same way, across our four target logics. But more could (and should) be said. Our unification sheds light on the logical and semantic similarities between the two kinds of paradoxes, but it does not tell us whether such similarities share a common origin. And, as of now, semantic and soritical paradoxes appear to rest on heterogeneous principles – naïveté and tolerance. Up to this point, our findings can be summarized in

[21] The proofs of Propositions 4.8-4.9, and 4.10-4.11 show that there are two senses of 'vacuity' at play. In the former case, the systems of equations associated with the premises of the argument is not solvable *in a way that makes them* s*-true*. In the latter case, the systems of equations are not solvable *at all*. The difference between these two senses of 'vacuity' matters: in the case of semantic and soritical paradoxes, Propositions 4.8 and 4.9 show that mne–validity (m=s) blocks the paradox, also via a vacuous case, in a way that is similar to what happens in a traditional, non-equational semantics, when a sentence cannot be made true in a class of models. On the other hand, in Proposition 4.10 and 4.11, the equational semantics shows that the sentences appearing in the argument *cannot be interpreted at all* by the target theories (i.e. assigned a value in the target value space). Therefore, 'vacuity' in this second sense is substantial, as it reveals the impossibility of blocking these paradoxes in the target semantics.

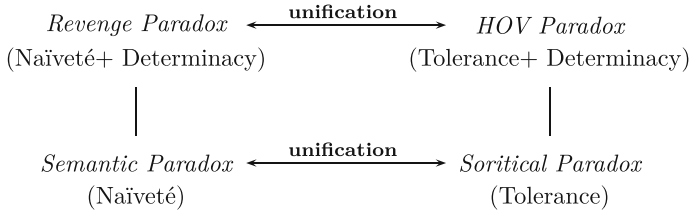

Fig. 5 Unification of paradoxes

Fig. 5: our unification allows us to connect the corresponding nodes of the tree, which haven't been shown to have a common root – yet.

However, since the semantic paradoxes and the soritical paradoxes behave remarkably similarly, supposing them to have a common root is rather natural. We now turn to the relations between tolerance and naïveté, and isolate a general schema which underlies them both. After some logical transformations, tolerance and naïveté will be shown to be nothing but *instances* of the same principle, embodying a notion of *indiscernibility*. We finally close the paper by hinting at some consequences for the solution of soritical and semantic paradoxes that indiscernibility suggests.

Here is the schema of indiscernibility that naïveté and tolerance will be reduced to:

$$[(\sim\text{- IND})]\ \forall x \forall y [x \sim y \rightarrow (\varphi(x) \leftrightarrow \varphi(y))],$$

where $\sim$ is a binary relation on terms, and $\varphi(x)$ is a schematic formula featuring at least one free occurrence of x, and possibly parameters.

In its purely schematic formulation, ($\sim$- IND) is not particularly informative: much of its plausibility depends on the relation replacing $\sim$, and the formulae replacing $\varphi(x)$. We can understand it as a schema, like the one considered in [74], providing intuitively necessary but not sufficient conditions for $\sim$ to express a similarity relation concerning instances of φ. Therefore, we now turn to specific classes of instances of ($\sim$- IND).

5.1 Tolerance and Naïveté as Instances of Indiscernibility

Our claim, simply put, is that both naïveté and tolerance can (and should) be understood as forms of indiscernibility. In order to defend this conclusion, we will show that:

(i) there are a relation $\sim_{\mathsf{To}}$ and a schematic formula $\varphi_{\mathsf{To}}(x)$ of $\mathcal{L}_{\mathrm{t,v}}$ s.t. tolerance follows deductively from the result of replacing $\sim_{\mathsf{To}}$ for $\sim$, and φ_{To} for φ;
(ii) there are a relation $\sim_{\mathsf{Tr}}$ and a formula φ_{Tr} of a *suitable extension of* $\mathcal{L}_{\mathrm{t,v}}$ s.t. naïveté follows deductively from the result of replacing $\sim_{\mathsf{Tr}}$ for $\sim$, and φ_{Tr} for φ.

Claim (i) is immediate. Recall (Section 1.3) that (TOLERANCE) is the following principle:

$$[(\text{TOLERANCE})]\ \forall x \forall y (\mathsf{P}(x) \wedge x \sim_{\mathsf{P}} y \rightarrow \mathsf{P}(y))$$

Then, one can prove (i) for every P by simply letting $\sim_{\mathsf{P}}$ to be $\sim_{\mathsf{To}}$ and $\mathsf{P}(x)$ to be $\varphi_{\mathsf{To}}(x)$. Claim (ii) is much less evident. First, it is not clear that a similarity relation

connected to the naïveté of truth can be found. Second, it is not clear that such a relation might allow us to deduce naïveté from indiscernibility. Let's address these two worries in turn.

First, appearances to the contrary, there seems to be a similarity relation implicitly at work in naïve truth. Let $\sim_{\mathsf{Tr}}$ be the relation over $\mathcal{L}_{\mathrm{t,v}}$-terms s.t. $s \sim_{\mathsf{Tr}} t$ holds if and only if either $s = \ulcorner \psi \urcorner$ and $t = \ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner$, or $t = \ulcorner \psi \urcorner$ and $s = \ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner$ for some formula ψ of $\mathcal{L}_{\mathrm{t,v}}$. Informally, $\sim_{\mathsf{Tr}}$ is a relation holding of codes of formulae just in case exactly one of the two is the truth predication of the other. Naïveté seems directly connected with this similarity relation, because it declares two formulae to be equivalent (or interderivable, or intersubstitutable) exactly when $\sim_{\mathsf{Tr}}$ holds between the terms denoting them.[22]

Let's now address the second problem highlighted above, and the claim (ii). Before providing the formal details, let us describe the argument informally.

In order to do this, we need a formula $\mathfrak{A}(x)$ of a language $\mathcal{L}_m \supseteq \mathcal{L}_{\mathrm{t,v}}$ such that, for s and t s.t. $s = \ulcorner \psi \urcorner$ and $t = \ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner$ ($\psi \in \mathcal{L}_{\mathrm{t,v}}$), $\mathfrak{A}(s)$ is equivalent to ψ and $\mathfrak{A}(t)$ is equivalent to $\mathsf{Tr}(\ulcorner \varphi \urcorner)$. In this way, we could formulate the following version of indiscernibility:

$$s \sim_{\mathsf{Tr}} t \rightarrow (\mathfrak{A}(s) \leftrightarrow \mathfrak{A}(t)) \tag{1}$$

Given the equivalences above, (1) can be read as saying that, if s and t code ψ and $\mathsf{Tr}(\ulcorner \psi \urcorner)$ respectively, then the corresponding instance of naïveté, i.e. that ψ is equivalent to $\mathsf{Tr}(\ulcorner \psi \urcorner)$), holds as well. In other words, (1) immediately enables one to derive naïveté (in the form of the (Tr- SCHEMA)), because it entails all the instances of the following schema:

$$\ulcorner \psi \urcorner \sim_{\mathsf{Tr}} \ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner \rightarrow (\psi \leftrightarrow \mathsf{Tr}(\ulcorner \psi \urcorner)) \tag{2}$$

as s and t are s.t. $s \sim_{\mathsf{Tr}} t$ vary. The following proposition makes the argument formally precise, by specifying a language $\mathcal{L}_m$ and a theory formulated in it where the required $\mathfrak{A}(x)$ can be constructed and proven to have the required property, thus enabling one to derive (2) from (1).

Proposition 5.1 *Let $\mathcal{L}^2_{\mathrm{t,v}}$ be the second order extension of $\mathcal{L}_{\mathrm{t,v}}$ that contains the language of second-order arithmetic. Let Δ^1_1-CA be the subsystem of second-order arithmetic with the comprehension axiom limited to Δ^1_1-formulae. Then, there is a set Γ of instances of ($\sim$- IND) formulated in $\mathcal{L}^2_{\mathrm{t,v}}$ s.t., for every $\psi \in \mathcal{L}_{\mathrm{t,v}}$:*

$$\Gamma \vdash_{\Delta^1_1\text{-CA}} \ulcorner \psi \urcorner \sim_{\mathsf{Tr}} \ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner \rightarrow (\psi \leftrightarrow \mathsf{Tr}(\ulcorner \psi \urcorner))$$

Proof Let $\mathsf{Sat}(x, y)$ be the formula expressing the fact that 'x is (the code of) a formula of $\mathcal{L}_{\mathrm{t,v}}$ satisfied by the term y', for the usual notion of Tarskian satisfiability. Recall that this is a Δ^1_1-definable relation (see, for instance, [71]), and hence its existence

[22] Notice that $\sim_{\mathsf{Tr}}$ is symmetric but not reflexive. If non-reflexivity appears undesirable for a similarity relation, one can simply stipulate that the relation holds also for identical terms.

is provable in Δ^1_1-CA. Let now $\mathsf{Tar}(x)$ be the formula $\forall y \mathsf{Sat}(x, y)$, i.e. the formula expressing that every term Tarski-satisfies the $\mathcal{L}_{\mathrm{t,v}}$-formula coded by x, i.e. that the formula is Tarski-true. Let Γ be the set of the following instances of ($\sim$- IND) in $\mathcal{L}^2_{\mathrm{t,v}}$, for all $\mathcal{L}_{\mathrm{t,v}}$-terms s and t:

$$s \sim_{\mathsf{Tr}} t \rightarrow (\mathsf{Tar}(s) \leftrightarrow \mathsf{Tar}(t)) \tag{3}$$

Recall that Δ^1_1-CA proves the disquotation schema restricted to $\mathcal{L}_{\mathrm{t,v}}$-formulae for Tar, i.e. for every $\chi \in \mathcal{L}_{\mathrm{t,v}}$:

[(DISQ)] $\mathsf{Tar}(\ulcorner \chi \urcorner) \leftrightarrow \chi$

(The proof, due to Tarski, is by induction on the complexity of χ). Let now s and t be $\mathcal{L}_{\mathrm{t,v}}$-terms s.t. $\vdash_{\Delta^1_1\text{-CA}} s = \ulcorner \psi \urcorner$ and $\vdash_{\Delta^1_1\text{-CA}} t = \ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner$. By (3):

$$\mathsf{Tar}(\ulcorner \psi \urcorner) \leftrightarrow \mathsf{Tar}(\ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner) \tag{4}$$

In turn, (4), and DISQ (together with some classical logic) imply:

$$\ulcorner \psi \urcorner \sim_{\mathsf{Tr}} \ulcorner \mathsf{Tr}(\ulcorner \psi \urcorner) \urcorner \rightarrow (\psi \leftrightarrow \mathsf{Tr}(\ulcorner \psi \urcorner)) \tag{5}$$

as desired. □

Claims (i) and (ii) allow us to conclude that both tolerance and naïveté are instances of ($\sim$- IND), in suitable languages. While the claim is obvious for tolerance, for naïveté it can be shown that, in suitable theories, one can formulate a version of indiscernibility that says that, if two $\mathcal{L}_{\mathrm{t,v}}$-terms code sentences s.t. one is the truth-predication of the other, then the two sentences are equivalent, and then derive all instances of the Tr-SCHEMA from such instances, in any theory that enables us to formulate these very instances of ($\sim$- IND).[23,24]

Some answers to possible objections are in order. First, one could object that, since the proof of Proposition 5.1 is carried out in classical logic, our reduction of naïveté to indiscernibility presupposes accepting classical logic. However, notice that the target reduction can be performed in ss, tt, ts, and st, by considering inferential or meta-inferential formulations of (3). Unsurprisingly, the forms of naïveté that result from

[23] The proof of Proposition 5.1 could be generalized and provide versions of the Tr-SCHEMA with bound variables by employing $\mathsf{Sat}(x, y)$ rather than $\mathsf{Tar}(x)$.

[24] One proposal to unify semantic, set-theoretic, and soritical paradoxes is Priest's *Inclosure Schema* [63, 65]. Comparing indiscernibility and inclosure is beyond the scope of this work, but the two approaches are evidently distinct. Indiscernibility is a schema from which tolerance and naïveté can be derived, while inclosure aims at modelling paradoxical reasonings. As such, the inclosure schema requires a specific set-theoretic and logical framework. Indiscernibility, by contrast, is designed to be independent from one's logic, so that paradoxical reasonings can be analyzed within the equational framework, which allows one to vary every component of the target theory – the base model, the interpretation of the logical constants, and the notion of logical consequence. These considerations suggest that our framework is more general and flexible.

such reformulation are exactly the version of the Tr-SCHEMA that is validated in sst, ttt, tst, and stt, detailed in Proposition 2.4.

One can object that the latter derivation is carried out in a second-order theory, formulated in a language that extends $\mathcal{L}_{t,v}$, and hence does not involve instances of ($\sim$- IND) formulated in $\mathcal{L}_{t,v}$, which shows that instances of the Tr-SCHEMA for $\mathcal{L}_{t,v}$ cannot be derived from indiscernibility. This objection is quite weak: schemata are not related to a specific language, but are typically understood as open-ended, so that their acceptance entails accepting their instances in richer languages [21]. If this is the case, then having instances of ($\sim$- IND) in $\mathcal{L}^2_{t,v}$ deriving instances of the Tr-SCHEMA for $\mathcal{L}_{t,v}$ in a relatively modest meta-theory does not invalidate our claim that all instances of the Tr-SCHEMA are derivable from ($\sim$- IND).

A stronger objection is this: our derivation of the Tr-SCHEMA from instances of ($\sim$- IND) is circular because it purports to derive the disquotational behaviour of the truth predicate using a disquotational truth predicate, that is $\mathsf{Tar}(x)$. The objection is reasonable but off-target. What we want to derive from ($\sim$- IND) – the Tr-SCHEMA or another form of naïveté – is a non-provable property of the object-linguistic truth predicate. What we use in deriving it – the schema DISQ – is a mathematical fact, that can be proven in any theory that defines the Tarskian truth predicate Tar, which we use instrumentally for this purpose. So, rather than postulating the naïveté of Tr to derive the naïveté of Tr (which would be circular), we employ provable mathematical facts (namely DISQ) about the Δ^1_1-definable $\mathcal{L}^2_{t,v}$-formula Tar to show that naïveté for Tr follows from ($\sim$- IND). To be sure, DISQ *resembles* naïveté, since it is essentially its typed restriction. But this does not make the two principles equivalent. By comparison, consider the set-theoretic principles of separation and naïve comprehension. They also resemble each other: the former is a consistent restriction of the latter. But this hardly makes them equivalent. While separation is a commonly accepted mathematical principle, naïve comprehension is not. And, to continue our analogy, any set-theoretic derivation that employs separation (and other principles) to show in some $T \supseteq \mathsf{ZF}$ that a certain schema entails naïve comprehension could not be suspected of circularity for its use of separation.

Finally, a word on what we take our reduction to do and not to do. We do not advance any empirical claim: we do not claim speakers to have in mind (if implicitly) anything like indiscernibility when they use vague or semantic predicates tolerantly or naïvely, nor is indiscerbibility taken from, or corroborated by empirical observation. Our reduction is not prescriptive either: we do not claim that speakers or theorists should understand vague or semantic notions via indiscernibility. Our reduction is a conceptual analysis, that shows how naïveté and tolerance derive from a unique, more fundamental notion and, by doing that, offers an explanation of their relationships. Such an analysis, in turn, explains the similarities between theories of truth and vagueness (and their treatment of paradoxes) which employ the same logics that were noted at the beginning of the paper, and that were made explicit in Section 4, via their equational re-formulation. In what follows, we briefly elaborate on the import of the connections between naïveté and tolerance that their analysis via indiscernibility reveals.

5.2 Paradoxes as Failure to see the Differences

Indiscernibility says that, if s and t are similar (in the sense of $\sim$), something (in the substitution class of φ) can be said of s just in case it can be said of t. If an application of indiscernibility leads to paradox,[25] it is reasonable to conclude that the paradox derives from having declared that s and t be similar (in the relevant respect). In other words, the paradox arises from applying the similarity relation too 'broadly', as in some cases such a similarity leads to contradictions. Tolerance and naïveté arise from a *failure to see a difference* between (i) two individuals that are similar to each other, as far as the vague predicate P is concerned, and (ii) (two terms coding) a sentence φ and its truth-predication $\mathsf{Tr}(\ulcorner\varphi\urcorner)$. And such a failure to differentiate, sometimes, leads to paradox. As seen in Sections 2 and 3, one might *try* to block the paradoxes by adopting a non-classical logic. But non-classical theories can only partially solve the problem: semantic and soritical paradoxes resurrect, and affect these very theories, in the form of revenge and HOV paradoxes. To our mind, the moral is this: the root of the paradoxes does not rely in not having yet identified a (or the) 'right' non-classical logic for truth and vagueness, but in the principle of indiscernibility itself, and in the failure to see the differences that flows from it.

Let us elaborate a bit on the idea of failure to see the differences. Which differences are relevant to discriminate between two individuals is, largely, a matter of *scale*. Accepting (i) and (ii), therefore, means that one has implicitly adopted a scale (of the relevant kind) that fails to discriminate between any two individuals (of the relevant kind). But which scales are (if implicitly) at work when it comes to tolerance and naïveté? Consider the vagueness case first. If P is a one-dimensional vague predicate, declaring whether s and t are P-similar depends on choosing a unit of measure for P's dimension. If P is the predicate 'tall', a difference of 1mm between two individuals does not seem relevant, whereas a difference of 10cm does. By our analysis, a soritical paradox involving 'tall' implicitly involves the adoption of a scale that is not fine-grained enough to discriminate between the heights of the individuals in question. In the case of truth, we suggest that the relevant difference between φ and $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ is to be cashed out in terms of *complexity*. By 'complexity', here, we do not mean logical complexity (φ can be of arbitrary logical complexity, whereas $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ is atomic), but the computational complexity of the definition of the truth predicate. Let us explain. The truth predicate can be seen under two main lights: as a 'simple' concept, or as a 'complex' one.[26] According to the first interpretation (mainly adopted by deflationists), the truth predicate is just a linguistic device that enables one to express generalizations. As such, the concept of truth is not especially complex: it is taken as a primitive notion, and given an axiomatization; therefore, its interpretation is not in need of a computationally complex model-theoretic definition [43]. The second interpretation postulates a significant difference between accepting φ and accepting $\mathsf{Tr}(\ulcorner\varphi\urcorner)$: accepting $\mathsf{Tr}(\ulcorner\varphi\urcorner)$ (the 'semantic ascent', as it is sometimes called) requires accepting all the resources employed in constructing an interpretation for the truth

[25] As in a soritical series, or declaring $\ulcorner\lambda\urcorner$ and $\ulcorner\mathsf{Tr}(\ulcorner\lambda\urcorner)\urcorner$ to be $\sim_{\mathsf{Tr}}$-similar and reproducing the Liar Paradox.

[26] For more details on the computational complexity of truth, see e.g. [7, 44, 48].

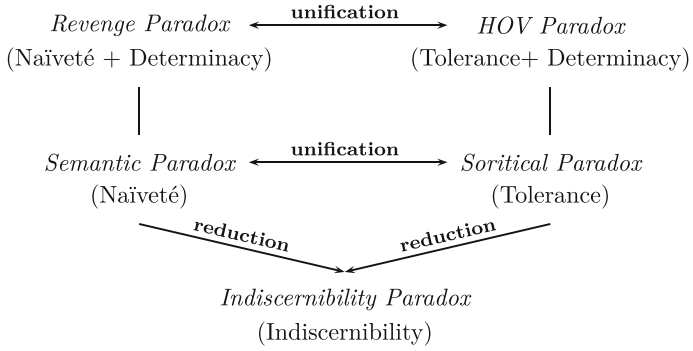

Fig. 6 Our reduction of paradoxes

predicate – Tarskian or Kripkean lines. And such resources are highly non-trivial, both from an epistemic point of view (the notions required for a Tarskian or Kripkean definition of truth are sophisticated logico-linguistic devices), and from the point of view of computational complexity.[27]

Going back to indiscernibility: if paradoxes arise from a failure to see a difference, then semantic paradoxes arise from failing to distinguish between the mere acceptance of a sentence φ and the acceptance of its truth-predication $\mathsf{Tr}(\ulcorner\varphi\urcorner)$. Advocates of the simple conception of truth are naturally led to accommodate the 'naïveté' intuition, according to which a sentence is simply equivalent to its truth-predication, and to look for a possibly chimeric revenge-free non-classical theory. But if failing to see a difference in complexity is what brings about the semantic paradoxes, then one should just reject the idea that truth is simple, accept that it is in fact a complex notion, and adopt a *hierarchical* interpretation of the notion of truth, that takes the complexity of truth-predication into account [30–33]. So, if the paradoxes do arise from a failure to see a difference in the relevant scale, then theories that postulate different standards for interpreting vague expressions and truth predications throughout the relevant scales – notably, *contextualist solutions* – have a better chance to solve the paradoxes at their root. It seems therefore possible to develop unified contextualist theories of truth and vagueness, that solve the indiscernibility paradoxes altogether.

To conclude, we can now articulate what the reduction of naïveté and tolerance to indiscernibility adds to our analysis. Given the 'horizontal' correspondence between the two kinds of paradoxes we obtained at the beginning of Section 5, we can now also uncover a 'vertical' connection which includes semantic and soritical paradoxes as examples of indiscernibility paradoxes. The combined result of unification and reduction is visualized in Fig. 6.

[27] The definition of the least Kripkean fixed point is Π^1_1-complete. For an overview of both interpretations of the truth predicate (roughly along the lines of the deflationism/inflationism debate), see [4, 53].

6 Conclusions

We have argued that semantic and soritical paradoxes are two sides of the same coin. Moreover, we have argued that the same holds, *mutatis mutandis*, for revenge and HOV paradoxes. To make our analysis more concrete, we focused on the Liar Paradox, the Sorites Paradox, and a family of three-valued logics. We introduced an equational framework to formalize them, and we showed that they are essentially similar pieces of reasoning. In addition, we have identified a schema of indiscernibility as the more general principle that underscores both naïveté and tolerance, tracing the Liar and the Sorites back to it. We finally argued that indiscernibility could be regarded as the unique source of both paradoxes.

Much work remains to be done. First, our analysis could be straightforwardly extended to more semantic paradoxes, including the paradoxes of denotation, satisfaction, naïve validity, and more. Second, other logical frameworks should be explored, to widen the scope of unification of paradoxes across the logical space: many-valued approaches with more than 3 values, super- and subvaluational approaches, and so on. Third, further kinds of paradoxes, e.g. involving set membership and property instantiation, could be explored.

Acknowledgements The authors would like to thank two anonymous referees for their useful comments over a previous version of the paper. They also wish to express their gratitude to Paul Égré and David Ripley, for commenting on an early version of the paper, and to Damian Szmuc for his feedback and his important suggestion to extend our unification proposal from revenge and higher-order vagueness to 'ordinary' semantic and soritical paradoxes. Finally, they would like to thank Kentaro Fujimoto, Simone Picenni, and Johannes Stern for a useful discussion concerning the reduction of the T-Schema to indiscernibility.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement. For Riccardo Bruni: this work was supported by the Italian Ministry of Education, University and Research through the PRIN 2017 program "The Manifest Image and the Scientific Image" prot. 2017ZNWW7F_004.
For Lorenzo Rossi: this work was supported by the Research Grant MSCA Staff Exchanges 2021 (HORIZON-MSCA-2021-SE-01) no. 1010866295 "PLEXUS: Philosophical, Logical, and Experimental Routes to Substructurality".

References

1. Alxatib, S., & Pelletier, F. (2011). The psychology of vagueness: Borderline cases and contradictions. *Mind & Language, 26*(3), 287–326.
2. Barrio, E., Rosenblatt, L., & Tajer, D. (2015). The logics of strict-tolerant logic. *Journal of Philosophical Logic, 44*(5), 551–571.
3. Beall, J. C. S. (2009). *Spandrels of Truth*. Oxford: Oxford University Press.

4. Beall, J. C., & Glanzberg, M. (2008). Where the paths meet: Remarks on truth and paradox. *Midwest Studies in Philosophy, 32*(1), 169–198.
5. Black, M. (1937). Vagueness. An exercise in logical analysis. *Philosophy of Science,4*(4):427–455.
6. Bochvar, D. A. (1937). [On a three-valued calculus and its applications to the paradoxes of the classical extended functional calculus]. Mathematicheskii sbornik, 4(46):287–308. (1937). English translation by M. *Bergmann in History and Philosophy of Logic, 2*(1981), 87–112.
7. Burgess, J. (1986). The truth is never simple. *Journal of Symbolic Logic, 51*(3), 663–681.
8. Burnett, H. (2017). *Gradability in natural language: Logical and grammatical foundations* (Vol. 7). Oxford: Oxford University Press.
9. Cantini, A. (1991). A theory of formal truth arithmetically equivalent to id1. *Journal of Symbolic Logic, 55*(1), 244–259.
10. Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2012). Tolerant, classical, strict. *Journal of Philosophical Logic, 41*(2), 347–85.
11. Cobreros, P., Égré, P., Ripley, D., & van Rooij, R. (2013). Reaching transparent truth. *Mind, 122*(488), 841–866.
12. Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2015). Vagueness, truth and permissive consequence. In T. Achourioti et al., (Eds.), *Unifying the Philosophy of Truth*, pages 409–430. Springer, 2015.
13. Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2017). Tolerant reasoning: nontransitive or nonmonotonic? *Synthese*, pages 1–25.
14. Davidson, D. (1967). Truth and meaning. *Synthese, 17*, 304–23.
15. Dummett, M. (1975). Wang's paradox. *Synthese, 30*, 301–324.
16. Égré, P. (2015). Vagueness: Why do we believe in tolerance? *Journal of Philosophical Logic, 44*(6), 663–679.
17. Égré, P. (2021). Half-Truths and the Liar. In C. Nicolai and J. Stern (Eds.), Modes of Truth. The Unified Approach to Modality, Truth, and Paradox. Routledge, London.
18. Égré, P., Ripley, D., & Verheyen, S. (2019). The sorites paradox in psychology. In S. Oms & E. Zardini (Eds.), *The Sorites Paradox.* Cambridge: Cambridge University Press.
19. Égré, P., et al. (2011). Perceptual ambiguity and the sorites. In R. Nouwen (Ed.), *Vagueness in Communication* (pp. 64–90). Springer.
20. Eklund, M. (2005). *What Vagueness Consists In. Philosophical Studies, 125*(1), 27–60.
21. Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic, 56*(1), 1–49.
22. Field, H. (2004). The semantic paradoxes and the paradoxes of vagueness. In J.c. Beall (Ed.), Liars and heaps: New essays on paradox. Oxford University Press.
23. Field, H. (2007). Solving the paradoxes, escaping revenge. In J. C. Beall (Ed.), Revenge of the Liar (pp. 53–144). Oxford University Press.
24. Field, H. (2008). *Saving Truth from Paradox*. Oxford: Oxford University Press.
25. Field, H. (2014). Naive truth and restricted quantification: Saving truth a whole lot better. *Review of Symbolic Logic, 7*(1), 147–191.
26. Fine, K. (1975). Vagueness, truth and logic. *Synthese, 30*, 265–300.
27. Gaifman, H. (2010). Vagueness, tolerance and contextual logic. *Synthese, 174*(1), 5–46.
28. Gillon, B. (1990). Ambiguity, generality, and indeterminacy: Tests and definitions. *Synthese, 85*(3), 391–416.
29. Gillon, B. (2004). Ambiguity, indeterminacy, deixis and vagueness: evidence and theory. In S. Davis & B. Gillon (Eds.), *Semantics: A Reader* (pp. 157–190). Oxford: Oxford University Press.
30. Glanzberg, M. (2001). The liar in context. *Philosophical Studies, 103*(3), 217–51.
31. Glanzberg, M. (2004). A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic, 33*, 27–88.
32. Glanzberg, M. (2004). Truth, reflection, and hierarchies. *Synthese, 142*(3), 289–315.
33. Glanzberg, M., et al. (2015). Complexity and Hierarchy in Truth Predicates. In T. Achourioti (Ed.), *Unifying the Philosophy of Truth.* Springer.
34. Glanzberg, M., & Rossi, L. (2000). Truth and quantification. unpublished manuscript.
35. Goguen, J. (1969). The logic of inexact concepts. *Synthese, 19*(3–4), 325–373.
36. Graff, D. (2000). Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics, 28*(1), 45–81.
37. Greenough, P. (2003). Vagueness: A minimal theory. *Mind, 112*(446), 235–281.
38. Halbach, V., & Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic, 71*, 677–712.

39. Halldén, S. (1949). The logic of nonsense. Uppsala University Press.
40. Hansen, N., & Chemla, E. (2017). Color adjectives, standards, and thresholds: An experimental investigation. *Linguistics and Philosophy, 40*(3), 239–278.
41. Heck, R. K. (2004). Semantic accounts of vagueness. In Jc Beall (Ed.), Liars and Heaps. New Essays on Paradox, pages 106–27. Oxford University Press, Oxford.
42. Heck, R. K. (2007). Self-reference and the languages of arithmetic. *Philosophia Mathematica, 15*(1), 1–29.
43. Horsten, L. (2012). The Tarskian Turn. Deflationism and axiomatic truth. MIT Press, Cambridge, (Mass.)
44. Horsten, L., & Leigh, G. (2017). Truth is simple. *Mind, 126*(501), 195–232.
45. Kamp, H. (1975). Two theories of adjectives. In E. Keenan (Ed.), *Formal semantics of natural language* (pp. 123–155). Cambridge: Cambridge University Press.
46. Kamp, H. (1981). The paradox of the heap. In U. Mönnich (Ed.), *Aspects of Philosophical Logic* (pp. 225–277). Dordrecht: Springer.
47. Kleene, S. C. (1952). *Introduction to Metamathematics*. New York: van Nostrand.
48. Kremer, P. (2009). Comparing fixed-point and revision theories of truth. *Journal of Philosophical Logic, 38*(4), 363–403.
49. Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72*, 690–716.
50. Lassiter, Dan, et al. (2011). Vagueness as probabilistic linguistic knowledge. In R. Nouwen (Ed.), *Vagueness in Communication* (pp. 127–150). Springer.
51. Machina, K. (1976). Truth, belief, and vagueness. *Journal of Philosophical Logic, 5*(1), 47–78.
52. McGee, V. (1991). *Truth, Vagueness, and Paradox*. Indianapolis: Hackett Publishing Company.
53. McGee, V. (2016). Thought, thoughts, and deflationism. *Philosophical Studies, 173*, 3153–3168.
54. McGee, V., & McLaughlin, B. (1995). Distinctions without a difference. *The Southern Journal of Philosophy, 33*(Supplement), 203–251.
55. Moschovakis, Y. (1974). *Elementary induction on abstract structures*. Amsterdam, London and New York: North- Holland and Elsevier.
56. Murzi, J., & Rossi, L. (2020). Generalised revenge. *Australasian Journal of Philosophy, 98*(1), 153–177.
57. Murzi, J., & Rossi, L. (2021). The expressive power of contextualist truth. In C. Nicolai and J. Stern, (Eds.), Modes of Truth. The Unified Approach to Modality, Truth, and Paradox. Routledge, London
58. Nicolai, C., & Rossi, L. (2018). Principles for object-linguistic validity: from logical to irreflexive. *Journal of Philosophical Logic, 47*, 549–577.
59. Nicolai, C., & Rossi, L. (). Truth. In P. Égré & L. Rossi (Eds.), Handbook of Trivalent Logic. Under contract with The MIT Press, Cambridge, Massachussetts, Forthcoming
60. Picollo, L. (2018). Reference in arithmetic. *The Review of Symbolic Logic*, pages 1–31.
61. Picollo, L., & Schindler, T. (2018). Disquotation and infinite conjunctions. *Erkenntnis, 83*(5), 899–928.
62. Priest, G. (1979). The logic of paradox. *Journal of Philosophical Logic, 8*, 219–241.
63. Priest, G. (2002). *Beyond the limits of thought*. Oxford: Oxford University Press.
64. Priest, G. (2006). *Doubt Truth to be a Liar*. Oxford: Oxford University Press.
65. Priest, G. (2010). Inclosures, vagueness, and self-reference. *Notre Dame Journal of Formal Logic, 51*(1), 69–84.
66. Ripley, D. (2012). Conservatively extending classical logic with transparent truth. *Review of Symbolic Logic*, pages 354–78.
67. Rossi, L. (2019). Model-theoretic semantics and revenge paradoxes. *Philosophical Studies, 176*(4), 1035–1054.
68. Rossi, L. (2019). A unified theory of truth and paradox. *Review of Symbolic Logic, 12*(2), 209–254.
69. Smith, N. J. J. (2008). *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.
70. Soames, S. (2004). Higher-order vagueness for partially defined predicates. In Jc. Beall (Ed.), Liars and heaps: New essays on paradox. Oxford: Oxford University Press.
71. Takeuti, G. (2013). *Proof Theory*. Mineola, New York: Dover Publication.
72. Tappenden, J. (1993). The liar and sorites paradoxes: Toward a unified treatment. *The Journal of Philosophy, 90*(11), 551–577.
73. Tennant, N. (2015). A new unified account of truth and paradox. *Mind, 124*(494), 571–605.
74. van Rooij, R. (2011). Vagueness and linguistics. In G. Ronzitti (Ed.), *Vagueness: A Guide*. Dordrecht: Springer.

75. Verheyen, S., Dewil, S., & Égré, P. (2018). Subjectivity in gradable adjectives: The case of tall and heavy. *Mind & Language, 33*(5), 460–479.
76. Verheyen, S., & Égré, P. (2018). Typicality and graded membership in dimensional adjectives. *Cognitive Science, 42*(7), 2250–2286.
77. Weir, A. (2005). Naïve truth and sophisticated logic. In JC Beall and B. Armour- Garb, (Eds.), *Deflationism and Paradox*, pages 218–249. Oxford University Press, Oxford
78. Wright, C. (1975). On the coherence of vague predicates. *Synthese, 30*, 325–365.
79. Zardini, E. (2008). A model of tolerance. *Studia Logica, 90*(3), 337–368.