Philos Stud (2019) 176:1035–1054
https://doi.org/10.1007/s11098-018-1041-7

Model-theoretic semantics and revenge paradoxes

Lorenzo Rossi[1]

Published online: 2 February 2018

Abstract Revenge arguments purport to show that any proposed solution to the semantic paradoxes generates new paradoxes that prove that solution to be inadequate. In this paper, I focus on revenge arguments that employ the model-theoretic semantics of a target theory and I argue, *contra* the current revenge-theoretic wisdom, that they can constitute genuine expressive limitations. I consider the anti-revenge strategy elaborated by Field (J Philos Log 32:139–177, 2003; Revenge of the Liar, Oxford University Press, Oxford, pp 53–144, 2007; Saving truth from paradox, Oxford University Press, Oxford, 2008, §§21–23) and argue that it does not offer a way out of the revenge problem. More generally, I argue that the difference between 'standard' and 'revenge' paradoxes is ill-conceived and should be abandoned. This will contribute to show that the theories that provide a uniform account of truth and other semantic notions are the ones best equipped to avoid the paradoxes altogether—'standard' and 'revenge' alike.

Keywords Model-theoretic semantics · Semantic paradoxes · Revenge paradoxes · Model-theoretic instrumentalism

1 Introduction

Prima facie, revenge paradoxes can be characterized as arguments to the effect that any proposed solution to the semantic paradoxes generates new paradoxes that prove that solution to be inadequate. Such paradoxes often make use of notions employed in the theories they are directed against, and are argued to be similar to

✉ Lorenzo Rossi
lorenzo.rossi@sbg.ac.at

[1] Department of Philosophy (KGW), University of Salzburg, Franziskanergasse 1, 5020 Salzburg, Austria

the 'standard' semantic paradoxes, such as the Liar and Curry's paradox (see e.g. Field 2007).[1] Revenge arguments are typically used to conclude that revenge-prone theories do not solve the semantic paradoxes in general: even though they avoid the 'standard' semantic paradoxes, they suffer from new, structurally similar antinomies, that can only be avoided at the cost of significant expressive limitations.

A straightforward revenge strategy involves the *model-theoretic* (henceforth 'MT') semantics of a theory of truth. MT-revenge arguments point to the inexpressibility in a theory T of some notion that is definable in or justified by the MT-semantics for T. For instance, let λ be a sentence equivalent to 'λ is not true', and let T be a non-classical theory of truth in which λ is not in the extension of the truth predicate, nor in the extension of the negated truth predicate.[2] In the MT-semantics for T, it is typically possible to define a predicate for 'determinateness' that captures the status of λ, declaring it to be not determinate. However, expressing such a predicate in T makes the theory T *trivial*—i.e. it forces T to contain every sentence of its language—and hence such a predicate is inexpressible in T. The inexpressibility of the model-theoretical notion of 'determinateness' provides an example of an MT-revenge paradox. As we will see below, the MT-revenge paradox just sketched can be adapted to essentially all non-classical theories of truth.[3]

The importance of MT-revenge derives from its scope. Most theories of truth come with an MT-semantics, which is used to provide an interpretation for them, or to prove them non-trivial. However, MT-revenge paradoxes are often considered to be as easy to construct as they are to defuse, and ultimately unproblematic.[4] In this paper, I challenge this view. In order to do so, I consider the vigorous defence against MT-revenge elaborated by Field (2007, 2008, §§21–23). Field's anti-revenge strategy can be applied to every theory for which an MT-semantics can be given: if successful, it promises to shield every theory of truth from revenge attacks based on MT-semantics. In a nutshell, Field argues that MT-revenge arguments depend on a confusion between *model-independent* and *model-relative* semantic notions: MT-revenge paradoxes employ the latter, but only the former can be used to characterize genuine semantic notions, and hence to give rise to genuine semantic paradoxes. However, I will argue that this distinction does not provide an acceptable ground to distinguish between genuine and non-genuine semantic notions, and thus it does not offer a way out of the revenge problem. More generally, I will argue that the difference between 'standard' and 'revenge' paradoxes is ill-conceived and should ultimately be abandoned. This will contribute

[1] For simplicity, I will only consider semantic notions formalized as first-order predicates and operators.

[2] In other words, neither the sentence 'λ is true' nor the sentence 'λ is not true' are in T. See for example the theories developed in Kripke (1975), Field (2003, 2008) and Halbach and Horsten (2006).

[3] Revenge paradoxes also exist for *classical* theories of truth (see e.g. Bacon 2015): my arguments for MT-revenge generalize to classical theories as well, but I will not explicitly discuss classical approaches, in the interest of space.

[4] See e.g. Beall (2007) and Field (2007, 2008).

to show that the theories that provide a *uniform* account of truth and other semantic notions are the ones best equipped to avoid the paradoxes altogether.[5]

The purpose of this paper is not to argue that *every* MT-revenge argument is successful. As Beall (2007) has warned, MT-revenge paradoxes can be very simple to formulate, and this can make the resulting arguments 'too easy', and ultimately uninteresting. In order to work, revenge arguments must have a proper justification, and their success can only be determined on a case-by-case basis. What I will show is that MT-revenge arguments do not fail *qua* model-theoretic: if they fail, they do so so for specific reasons, such as an unconvincing justification or lack of importance.

The paper is structured as follows. In Sect. 2 I introduce some representative 'standard' and 'revenge' semantic paradoxes. In Sect. 3 I present Field's anti-revenge strategy, and in Sect. 4 I argue that it fails to neutralize MT-revenge arguments. In Sect. 5 I discuss a possible rejoinder, namely the idea that the MT-semantics is a mere tool to prove non-triviality results and, as such, it does not characterize genuine semantic notions; I argue that this reply also fails to neutralize MT-revenge arguments. Section 6 concludes.

2 Paradoxes and revenge

Semantic paradoxes arise from a combination of three factors: (i) classical logic;[6] (ii) a *modicum* of syntax; (iii) *naïveté* about truth, namely the idea that for every sentence φ, φ and 'φ is true' are (in some sense) equivalent. To see this, consider a first-order language $\mathcal{L}_{\mathsf{Tr}}$ that contains the predicate True (for '... is true'), and a theory T formulated in $\mathcal{L}_{\mathsf{Tr}}$ that obeys classical logic (factor (i)) and is able to define a function $\ulcorner \cdot \urcorner$ that assigns names to sentences (factor (ii)). Let now suppose that T encodes some form of naïveté (factor (iii)); more specifically, let us assume that T contains all the instances of the following schema:

(T-SCHEMA) $\quad \mathsf{True}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi,$

or that T is closed under an *inter-substitutivity* rule, by which

(INTER-SUBSTITUTIVITY) $\quad \varphi \in T$ if and only if $\varphi^{\mathsf{t}} \in T,$

where φ^{t} is the result of substituting a subformula ψ of φ with $\mathsf{True}(\ulcorner \psi \urcorner)$ or *vice versa*. Finally, suppose that T can prove the existence of a sentence λ that is

[5] My focus in this paper is on MT-revenge and on Field's treatment thereof. Several general treatments of revenge paradoxes can be found in the literature, see e.g. Beall (2006, 2007, 2007b), Cook (2007), Eklund (2007), Maudlin (2007), Priest (2007), Restall (2007), Scharp (2007, 2013), Simmons (2007, Shapiro 2011 and Scharp (2013). In particular, Ketland (2003) and Beall (2007) discuss MT-revenge paradoxes in connection with classical meta-theories, and Leitgeb (2007) focuses on revenge paradoxes affecting the theory developed in Field (2003, 2007, 2008) in classical and non-classical meta-theories. In this paper, I focus on Field's argument, since it is meant to apply to every MT-semantics, irrespective of whether they are defined in a classical or non-classical meta-theory, and it applies to every theory of truth that has been given an MT-semantics.

[6] Actually, semantic paradoxes arise also in sufficiently strong, but non-classical, logics, such as intuitionistic logic. I will focus on classical logic for simplicity.

equivalent to $\neg\mathsf{True}(\ulcorner\lambda\urcorner)$ in T—the existence of sentences such as λ can be proven in any theory that interprets a *modicum* of syntax. If T is non-trivial, there is a classical evaluation function v assigning the semantic value **1** to the sentences of T. What, then, is the value of λ? It is easily seen that λ cannot be classically evaluated, on pain of contradiction. Since v is a classical evaluation, either $v(\lambda) = \mathbf{1}$ or $v(\lambda) = \mathbf{0}$. If $v(\lambda) = \mathbf{1}$, then $v(\neg\mathsf{True}(\ulcorner\lambda\urcorner)) = \mathbf{1}$ (by definition of λ), but also $v(\neg\lambda) = \mathbf{1}$ (courtesy of naïveté), which is absurd. We conclude that $v(\lambda) = \mathbf{0}$, and therefore $v(\neg\mathsf{True}(\ulcorner\lambda\urcorner)) = \mathbf{0}$ (by definition of λ). But the latter, by naïveté, yields $v(\neg\lambda) = \mathbf{0}$, which is also absurd. This is the Liar paradox.[7]

There are several options to restrict classical logic in order to non-trivially admit some form of naïveté. Semantically, this corresponds to adopting non-classical evaluation functions. Classical evaluations assign to every sentence either value **1** or value **0**, and no sentence is assigned the same classical value as its negation. This sits poorly with naïveté: the Liar paradox features a sentence λ that is forced to have the same value as its negation. However, several non-classical evaluations feature three or more semantic values: they behave as classical evaluations on classical values, but λ and $\neg\lambda$ can be assigned the same non-classical value. In this way, non-classical evaluations can assign the same value to φ and $\mathsf{True}(\ulcorner\varphi\urcorner)$ (and thus validate INTER-SUBSTITUTIVITY or even the T-SCHEMA).[8]

In order to make the following discussion more precise, I will now introduce a family of non-classical logics that can be used to formulate several theories of naïve truth. Moreover, I will recall a few basic facts about one specific theory of naïve truth, the one developed in Field (2003, 2007, 2008), which will make it easier to discuss Field's anti-revenge strategy. However, both Field's anti-revenge strategy and my arguments against it are completely independent from the revenge-breeding notions and the specific theories being considered.

Let a *partial evaluation* be any function that assigns to the sentence of $\mathcal{L}_{\mathsf{Tr}}$ one of the values **1**, **0**, and **1/2**, and that satisfies the following criteria:

The value of $\neg\varphi$ is **1** *minus* the value of φ.
The value of $\varphi \wedge \psi$ is the *minimum* of the values of φ and ψ.
The value of $\forall x\varphi$ is the *infimum* of the values of its instances $\varphi(t)$.

The other logical constants are defined as usual (and evaluated accordingly): $\varphi \vee \psi$ is $\neg(\neg\varphi \wedge \neg\psi)$, $\varphi \rightarrow \psi$ is $\neg(\varphi \wedge \neg\psi)$, and $\exists x\varphi$ is $\neg\forall x\neg\varphi$. Several non-classical logics that support some form of naïveté are based on partial evaluations. Strong Kleene logic, or $\mathsf{K3}$, is a case in point: a sentence φ is a $\mathsf{K3}$-consequence of a set of

[7] In keeping with the model-theoretic focus of this paper, I present the Liar and other paradoxes semantically, and I only consider theories of truth for which an MT-semantics has been developed. While these restrictions rule out theories that have only been developed axiomatically (e.g. Zardini 2011), nothing prevents the arguments in this paper from applying to MT-semantics that have yet to be developed.

[8] Several many-valued logics can be given two-valued semantics: see Chemla et al. (2017) for a systematic study and Rosenblatt (2015) for applications to theories of truth. However, MT-revenge paradoxes can be reproduced in such alternative two-valued frameworks, and therefore I will not explicitly consider them.

sentences Γ if, for every partial evaluation p, if $p(\Gamma) = \mathbf{1}$, then $p(\varphi) = \mathbf{1}$ (where $p(\Gamma) = \mathbf{1}$ is a shorthand for $p(\psi) = \mathbf{1}$, for every ψ in Γ). Using K3, one can give non-trivial theories of truth that validate INTER-SUBSTITUTIVITY while avoiding the truth-theoretical paradoxes.[9]

K3 is a very weak logic: on the one hand, several classically valid inference rules turn out to be invalid in it (notably, the classical rules for introducing negation and the conditional); on the other, K3 does not have any logical laws: not even the principle $\varphi \rightarrow \varphi$ is K3-valid.[10] Several theories have been developed that strengthen K3 without losing INTER-SUBSTITUTIVITY. In particular, a series of theories developed in recent years by Field (2002, 2003, 2008, 2013) succeeded in equipping K3-based theories of truth with primitive, strong conditional connectives, not equivalent to K3's conditional, that validate several classically valid principles.

Field's (2003, 2007, 2008) theory, call it F, is a case in point: it extends K3 with a primitive, strong conditional $\rightarrow_{\mathsf{F}}$, it contains all the instances of several classically valid schemata, such as $\varphi \rightarrow_{\mathsf{F}} \varphi$, $(\varphi \wedge \psi) \rightarrow_{\mathsf{F}} \varphi$, $\varphi \rightarrow_{\mathsf{F}} (\varphi \vee \psi)$, and it satisfies INTER-SUBSTITUTIVITY and the T-SCHEMA, where the latter is formulated with Field's biconditional, i.e. $\mathsf{True}(\ulcorner \varphi \urcorner) \leftrightarrow_{\mathsf{F}} \varphi$.[11] The set F is defined model-theoretically, namely via some evaluation function $v_{\mathscr{F}}$ from the sentences of the language to a set of semantic values (containing the classical values **1** and **0**):

$$\mathsf{F} = \{\varphi \in \mathsf{SENT} \mid v_{\mathscr{F}}(\varphi) = \mathbf{1}\},$$

where SENT indicates the set of sentences of the language of Field's theory. The evaluation functions employed by Field to define F assign a non-classical value to sentences such as λ, thus making it possible to non-trivially retain both INTER-SUBSTITUTIVITY and the T-SCHEMA. However, F is a very expressive theory, and it possesses the resources to characterize a *determinateness operator* Det that captures the status of λ and similar sentences.[12] Field's determinateness operator obeys the following rules (see Field 2007, pp. 110)[13]:

[9] For more on K3, see Kleene (1952) and Blamey (2002). For theories of naïve truth employing K3 (or closely related logics), see Kripke (1975), Kremer (1988), Halbach and Horsten (2006) and Horsten (2009).

[10] The reason for this is easily stated: while **1** is the only designated value of K3, any formula whose subformulae are assigned value **1/2** by a partial evaluation p is itself assigned value **1/2** by p, so no formula receives value **1** under all partial evaluations. For example, if $p(\varphi) =$**1/2**, then $p(\varphi \rightarrow \varphi) =$ **1/2** as well.

[11] See Field (2003, §4) and Field (2008, Chapter17.4) for a partial list of the principles validated in Field's theory. See Rossi (2016) for a self-contained presentation of Field's theory, and a discussion of its conditional.

[12] For simplicity, the notion of determinateness is treated as an operator Det, namely as a syntactic connective applying to sentences. A determinateness *predicate*, applying to names of sentences, is however easily definable in F, because F satisfies INTER-SUBSTITUTIVITY. So, a determinateness predicate DT obeying exactly the same principles of the operator Det can be defined putting $\mathsf{DT}(\ulcorner \varphi \urcorner) := \mathsf{True}(\ulcorner \mathsf{Det}(\varphi) \urcorner)$.

[13] As Field notes, D3 might be strengthened to the following condition: 'From $\varphi \rightarrow_{\mathsf{F}} \mathsf{Det}(\varphi)$, infer $\varphi \vee \neg\varphi$'. I will ignore this possibility for the sake of simplicity.

From $\varphi \rightarrow_{\mathsf{F}} \psi$, infer $\mathsf{Det}(\varphi) \rightarrow_{\mathsf{F}} \mathsf{Det}(\psi)$ (D1)
From φ, infer $\mathsf{Det}(\varphi)$ (D2)
$\mathsf{Det}(\varphi) \rightarrow_{\mathsf{F}} \varphi$ (D3)
From $\varphi \rightarrow_{\mathsf{F}} \neg\varphi$, infer $\neg\mathsf{Det}(\varphi)$ (D4)

Thanks to its determinateness operator, Field's theory can declare Liar sentences such as λ as 'not determinately true' – indeed, the sentence $\neg\mathsf{Det}(\mathsf{True}(\ulcorner\lambda\urcorner))$ is in F. Field's determinateness operator clearly allows one to form Liar-like sentences employing Det itself. The sentence λ^* inter-substitutable with $\neg\mathsf{True}(\ulcorner\mathsf{Det}(\lambda^*)\urcorner)$ in F is a case in point. Of course, the 'indeterminate' status of λ^* cannot be captured by declaring it 'not determinately true', since $\neg\mathsf{True}(\ulcorner\mathsf{Det}(\lambda^*)\urcorner)$ is inter-substitutable with λ^* itself. However, Field's theory declares λ^* to be 'not determinately determinately true', and in fact $\neg\mathsf{Det}(\mathsf{Det}(\mathsf{True}(\ulcorner\lambda\urcorner^*)))$ is in F. The iterations of Det definable in F go further, extending well into the transfinite (Field 2008, §§21–23).

F provides a solution to the Liar and all the other paradoxes that can be formulated in its language: not only does it validate INTER-SUBSTITUTIVITY and a form of the T-SCHEMA, it also provides a treatment of intuitively paradoxical sentences such as λ via (iterations of) the determinateness operator. Does this show that F, and similarly expressive theories of naïve truth more generally, solve *all* the semantic paradoxes? Revenge arguments aim to answer negatively to this question. More precisely, MT-revenge arguments aim at showing that theories of naïve truth, despite solving the truth-theoretic paradoxes, fall short of solving other semantic paradoxes closely related to the 'standard' ones, involving other semantic notions closely related to naïve truth, definable in their MT-semantics. The upshot of revenge arguments is clear: revenge-prone theories suffer from crucial *expressive limitations*: they avoid triviality only because they cannot express the semantic notions that could trivialize them, just like classical theories cannot express naïve truth.

Here is a classic MT-revenge paradox (see e.g. Ketland 2003; Leitgeb 2007)—I present it for an unspecified theory T, but one could easily run it for F:

Bivalent Determinateness Let T be a theory of truth that validates INTER-SUBSTITUTIVITY (but a similar argument applies for the T-SCHEMA), and let v be a non-classical evaluation for T. T cannot contain any operator BDet with the following semantics:

$$v(\mathsf{BDet}(\varphi)) = \begin{cases} \mathbf{1}, & \text{if } v(\varphi) = \mathbf{1}, \\ \mathbf{0}, & \text{if } v(\varphi) \neq \mathbf{1} \end{cases}$$

To see this, let λ_{d} be inter-substitutable for $\neg\mathsf{True}(\ulcorner\mathsf{BDet}(\lambda_{\mathsf{d}})\urcorner)$ in T, and apply v to λ_{d}:

- Suppose $v(\lambda_{\mathsf{d}}) = \mathbf{1}$. By definition of BDet and naïveté, $v(\mathsf{BDet}(\lambda_{\mathsf{d}})) = \mathbf{1} = v(\mathsf{True}(\ulcorner\mathsf{BDet}(\lambda_{\mathsf{d}})\urcorner))$. No evaluation assigns the same *classical* value to a sentence and its negation, but λ_{d} is by definition inter-substitutable for $\neg\mathsf{True}(\ulcorner\mathsf{BDet}(\lambda_{\mathsf{d}})\urcorner)$, and therefore $v(\neg\mathsf{True}(\ulcorner\mathsf{BDet}(\lambda_{\mathsf{d}})\urcorner)) \neq \mathbf{1} \neq v(\lambda_{\mathsf{d}})$, against our supposition.

– Suppose $v(\lambda_{\mathsf{d}}) \neq \mathbf{1}$. By definition of BDet and naïveté, $v(\mathsf{BDet}(\lambda_{\mathsf{d}})) = \mathbf{0} = v(\mathsf{True}(\ulcorner \mathsf{BDet}(\lambda_{\mathsf{d}}) \urcorner))$. Non-classical evaluations are classical on classical values, and λ_{d} is inter-substitutable for $\neg\mathsf{True}(\ulcorner \mathsf{BDet}(\lambda_{\mathsf{d}}) \urcorner)$, therefore $v(\neg\mathsf{True}(\ulcorner \mathsf{BDet}(\lambda_{\mathsf{d}}) \urcorner)) = \mathbf{1} = v(\lambda_{\mathsf{d}})$, against our supposition.

Whether λ_{d} is assigned $\mathbf{1}$ or a different value, λ_{d} and $\neg\mathsf{True}(\ulcorner \mathsf{BDet}(\lambda_{\mathsf{d}}) \urcorner)$ are impossibly forced to have the same semantic value, since BDet works as a 'classicalizer' for non-classical evaluations. Since T and v are arbitrary, no theory of naïve truth can feature an operator for bivalent determinateness.

Non-classical theorists typically deny the legitimacy of the above paradox, and relevantly similar ones, arguing that revenge-breeding notions such as bivalent determinateness are not genuine semantic notions.[14] In the next section, I will discuss an anti-MT-revenge strategy elaborated by Hartry Field in a series of works (2003, 2007, 2008). Even though Field's strategy is articulated in the context of his theory F, it can be applied to every theory for which an MT-semantics can be given.[15]

3 Field on MT-revenge

Field views MT-revenge arguments as the result of a confusion between *model-relative* and *model-independent* notions. Model-relative notions are essentially defined via reference to some model or evaluation function: 'having semantic value x' is a case in point. By contrast, truth is model-independent: it is characterized by principles such as the T-SCHEMA or INTER-SUBSTITUTIVITY that involve no model-theoretic reference. Field argues that genuine semantic notions are model-independent, and insofar as MT-revenge paradoxes resort to model-relative notions, they are not genuine paradoxes.

[14] The legitimacy of MT-revenge paradoxes cannot be questioned on purely formal grounds: BDet is definable in minimally strong meta-theories (Leitgeb 2007). One could object that BDet is definable only because a *classical* meta-theory is employed (Yablo 2003; Leitgeb 2007); a suitable non-classical meta-theory would make it undefinable. However, too little is known about (suitable) non-classical meta-theories to tell whether they would make *all* revenge-breeding notions undefinable. In Sect. 5 I present an argument for MT-revenge that does not rely on the classicality of one's meta-theory. For a non-classical meta-theory for a relatively weak object-theory, see Bacon (2013); for recent investigations on non-classical meta-theories, see Field et al. (2017).

[15] Using bivalent determinateness as a representative revenge paradox in the context of Field's theory may seem inappropriate, especially in view of verdicts such as the following one: '*there can be no truth-like predicate for which excluded middle can be assumed*' (Field 2007, p. 89, emphasis in the original). Nevertheless, Field acknowledges that 'the conviction that there *must* be truth-like predicates obeying excluded middle is one primary source of revenge worries' (*ibidem*). Likewise, he concedes that an argument against his theory based on bivalent determinateness 'is perhaps the one with most intuitive force: it is that we just need a unified [(i.e. bivalent)] account of determinacy or defectiveness' (Field 2008, p. 140)—I thank an anonymous referee for pointing out the latter quote to me. In any case, in the context of Field's theory, paradoxes structurally similar to the paradox of bivalent determinateness can be given without resorting to bivalence, e.g. using idempotent determinateness.

Field brings the following case in support of his claim. Consider the language of set theory expanded with a predicate True for '... is true'. A 'highly natural' model for this language is 'the [classical] homophonic model whose domain consists of all non-sets together with all sets of rank less than the first inaccessible cardinal'.[16] Call this model $\mathscr{M}_1$. In order to build $\mathscr{M}_1$, the existence of at least one inaccessible cardinal is required.[17] However, this is where model-independent truth and its model-relative counterpart diverge:

> But now consider the sentence 'There are inaccessible cardinals': it's true, but false in $\mathscr{M}_1$, i.e. has semantic value **0** in $\mathscr{M}_1$; its negation is false, but has value **1** in $\mathscr{M}_1$. Having semantic value **1** in $\mathscr{M}_1$ doesn't correspond to truth, or to determinate truth, or anything like that [...]. The point made here for $\mathscr{M}_1$ applies to any other model that can be defined within set theory, by Tarski's Theorem, and this includes all models of set theory that are at all 'natural'. (Field 2007, p. 104)

Since model-relative notions are relativized to some models, they always capture incorrectly the *extension* of model-independent notions. Models have sets as domains. As a consequence, a model-relative semantic notion N, defined relatively to a model $\mathscr{M}$ whose domain is M, can only have a subset of M as its extension. By contrast, model-independent notions are not relative to a particular model, and therefore their extension is not restricted to any particular set. This is exactly the divergence in extension between model-relative and model-independent truth that Field's quote points to. And it is because of such a divergence, Field argues, that the understanding of model-independent notions is not mediated nor conveyed by MT-definitions: one cannot 'extrapolate an understanding of a model-independent notion like truth or determinate truth from the model-relative notions' (Field 2007, p.105). But genuine semantic notions are not model-theoretically restricted, and are therefore model-independent, or so Field's argument suggests.

Field's argument purports to show that model-relative semantic notions cannot be genuine semantic notions: since they are always restricted to a set, they cannot correctly capture the extension of the notion (truth, determinateness, or else) they intend to characterize. Yet, Field's argument needs to apply even beyond model-relative notions: it needs to apply to model-independent semantic notions that are *motivated* or *justified* by the MT-semantics. As Field himself points out, revenge-breeding notions can in fact be characterized model-independently. Bivalent determinateness itself is a case in point: it is possible to (partially) *axiomatize* Det in such a way that it has to obey the evaluation clauses employed in the revenge paradox of Bivalent Determinateness. I will articulate this point in the context of Field's theory. Recall that in Field's theory F it is possible to define a

[16] Field (2007, p. 104). Homophonic models 'assign to a name its real bearer and analogously for function symbols, and [...] in the classical case assign to a predicate those objects in its real extension *that are also in the domain of the model*.' (Field 2007, *ibidem*) Field assumes for the sake of the example that there are inaccessible cardinals. If there are no inaccessible cardinals, the example can be altered accordingly.

[17] This is a basic fact of Zermelo-Fraenkel Set Theory. See for example Jech (2002, pp. 167–168).

determinateness operator Det that *approximates* bivalent determinateness, in that it obeys the clauses D1–D4 introduced on page 5. Crucially, Field's operator Det does not obey the Law of Excluded Middle:

$$\mathsf{Det}(\ulcorner\varphi\urcorner) \vee \neg\mathsf{Det}(\ulcorner\varphi\urcorner) \qquad (\mathrm{D5})$$

Adding D5 to D1–D4 would force the operator Det to behave just like the revenge-paradoxical operator BDet, namely it would turn Field's determinateness into a (partial) axiomatization of bivalent determinateness, thus trivializing the resulting theory.

Field is obviously aware of the fact that MT-revenge notions can be given model-independent formulations. Nonetheless, he argues that the resulting notions are not sufficiently well-motivated, and fail to yield genuine paradoxes.

> The proponent of the [MT-]revenge problem doesn't intend [bivalent determinateness] to be understood as model-relative. The question then arises, how is it to be understood. I do not deny that it is possible to introduce into the language an operator [...] with many of the features that the proponent of revenge wants [i.e. D1–D4], and which is *not* model-relative. [...] But such [operators] only breed paradox if they satisfy all the assumptions used in the [revenge-paradoxical] derivations [...]; the one place they fail is that excluded middle [i.e. D5] cannot be assumed for them. So there is a revenge problem [...] only if there is reason to think that we can understand a notion of [bivalent determinateness] that obeys those other assumptions *plus excluded middle*.
>
> And why assume that? I think what underlies the [MT-]revenge problem is the thought that the model-relative [bivalent determinateness operators] all obey excluded middle, so there must be an absolute [operator] that does too. But this assumption seems to me completely unwarranted: one just can't assume that one can extrapolate in this way from the case of model-relative predicates, which make sense only by virtue of 'misinterpreting' the quantifiers as having restricted range, to the unrelativized case where no such 'misinterpretation' is in force. (Field 2007, pp. 108–109)

The model-independent notion of bivalent determinateness (obeying D1–D5), Field argues, is only motivated by his model-relative counterparts: but how can the latter justify the former, given that model-relative notions fall inevitably short of determining the extension of the model-independent ones?

Summing up, Field's argument is based on the misalignment between model-independent and model-relative notions: the latter always fail to capture the extension of the former. However, model-relative notions can be given model-independent (revenge-breeding) formulations. Therefore, if Field's strategy is to succeed, his argument must apply to model-independent notions as well. For this reason, Field's argument needs some principle that links a sufficient understanding of a model-independent notion with its extension. Here's a first stab at such a bridge principle:

(EXTENSION) A sufficiently clear understanding of a model-independent notion N entails the existence of a criterion to determine the extension of N.

The need for a criterion to determine N's extension is suggested by Field's argument itself: one can be agnostic about the existence of inaccessible cardinals, but a sufficiently clear understanding of naïve truth requires the sentence 'There are inaccessible cardinals' to be in the extension of True just if there are large cardinals. Understanding truth, therefore, requires a suitable criterion to determine the extension of the corresponding notion.[18] If one wants to endorse Field's anti-revenge strategy, she has to endorse something like EXTENSION. Not accepting EXTENSION would amount to not accepting the only ground to discriminate between naïve truth and model-relative notions that Field's argument offers, namely that model-relative notions lack an extension which is not incorrectly determined.

Field's argument does not rely on the specificities of his theory, or of bivalent determinateness: if successful, it would defuse *every* MT-revenge argument. However, in the next section I'll argue that Field's argument must ultimately be rejected: MT-semantic notions, with their potential for revenge, still stand.[19]

4 Intelligibility and extensions

EXTENSION effectively extends Field's argument to model-independent notions motivated by the MT-semantics. However, it is too weak. The schemata D1–D5 that characterize bivalent determinateness model-independently provide criteria to determine its extension just like the T-SCHEMA or INTER-SUBSTITUTIVITY does for truth. Therefore, EXTENSION does not suffice to conclude that naïve truth is a genuine semantic notion, while D1-D5-determinateness is not, which is what Field's strategy aims to accomplish.

More precisely, the T-SCHEMA provides a criterion to determine the extension of True, according to which:

– '$2 + 2 = 4$' is in the extension of True if and only if $2 + 2 = 4$,

[18] There are several options for such a criterion, e.g. Dummett's criterion of applicability (Dummett 1993, pp. 75–77 and pp. 232–235). EXTENSION does not require a criterion that is effectively applicable, nor a criterion that determines completely the extension of a given notion. Field's argument only requires the extension of a notion N to be determined *not incorrectly*: the sentence 'there are large cardinals' should not fall outside of the extension of True if there are large cardinals.

[19] The relevance of Field's argument goes well beyond revenge paradoxes: if it were successful, it would have a profound impact on most truth-conditional semantic theories. On the one hand, Field's argument tells against any approach to semantics that relies heavily on model-theoretical resources, including several theories in the Montagovian tradition (see e.g. Chierchia and McConnell-Ginet (2000); for an argument against Montagovian semantics based on their use of model-relative notions, see Lepore (1983); for a recent reply, see Glanzberg (2014)). On the other, Field's argument tells against several theories in the tradition of Davidson (1967), which are based on model-independent notions motivated or justified by the MT-semantics (see e.g. Fischer et al. (2015, §3.1), Halbach (2014, Chapter 8)). Given its scope, one might worry that Field's anti-revenge argument rules out every approach to truth-conditional semantics. However, the deflationary approach to truth-conditions articulated in Field (1994) might be used to answer to this worry. See also Hill (2014, Part I) and McGee (2016) for a recent discussion.

– 'there are inaccessible cardinals' is in the extension of True if and only if there are inaccessible cardinals,

But D1–D5 can also be read as criteria to determine, perhaps partially but at least not incorrectly, the extension of Det:[20]

D1 - From a derivation of 'if $2 + 2 = 4$ then $3 + 2 = 5$', infer that if '$2 + 2 = 4$' is in the extension of Det then '$3 + 2 = 5$' is in the extension of Det,
- From a derivation of 'if there are inaccessible cardinals then there are large cardinals', infer that if 'there are inaccessible cardinals' is in the extension of Det then 'there are large cardinals' is in the extension of Det,

D2 - From a derivation of '$2 + 2 = 4$', infer that '$2 + 2 = 4$' is in the extension of Det,
- From a derivation of 'there are inaccessible cardinals', infer that 'there are inaccessible cardinals' is in the extension of Det,

D3 - if '$2 + 2 = 4$' is in the extension of Det, then $2 + 2 = 4$,
- if 'there are inaccessible cardinals' is in the extension of Det, then there are inaccessible cardinals,

D4 - From a derivation of 'if $2 + 2 = 4$ then it is not the case that $2 + 2 = 4$', infer that '$2 + 2 = 4$' is in the extension of the negated Det,
- From a derivation of 'if there are inaccessible cardinals then there are not inaccessible cardinals', infer that 'there are inaccessible cardinals' is in the extension of the negated Det,

D5 - Either '$2 + 2 = 4$' is in the extension of Det or '$2 + 2 = 4$' is in the extension of the negated Det,
- Either 'there are inaccessible cardinals' is in the extension of Det or 'there are inaccessible cardinals' is in the extension of the negated Det,

Summing up, EXTENSION is too weak to salvage Field's anti-revenge strategy because model-independent bivalent determinateness, i.e. Det axiomatized by D1–D5, *also* satisfies EXTENSION. D1–D5 provide criteria to determine the extension of bivalent determinateness in Field's theory—even though they make it trivial—and therefore EXTENSION fails to exclude bivalent determinateness from the genuine semantic notions.

In order to apply Field's anti-revenge strategy, and deny that D1–D5-bivalent determinateness is a genuine semantic notion, one has to exclude trivial concepts from the application of EXTENSION. This is the only option to salvage Field's argument, because naïve truth and bivalent determinateness (a) are both formulated in purely model-independent terms, and (b) both enjoy criteria to determine their

[20] Standardly, predicates have extension and operators don't. However, I'm ignoring this fact, given that in the presence of INTER-SUBSTITUTIVITY one can equally treat bivalent determinateness, or other semantic notions, as predicates or operators. See footnote 12.

extension (not incorrectly). For this reason, Field's argument really needs to be supplemented with the following principle:

(C-EXTENSION) A sufficiently clear understanding of a model-independent notion N entails the existence of a criterion that determines the extension of N, and it requires N's extension to be non-trivial.

C-EXTENSION is a very strong principle: it entails that if a notion N breeds paradox for a theory T, than N is not understandable for anyone accepting T. This seems deeply problematic, for it turns any attempt to compare theories into a futile exercise: the advocates of T would declare every notion incompatible with T to be simply nonsense. Several fundamental debates in theories of truth—including the debates on which is the 'right' non-classical logic of naïve truth, and the debates on whether truth is naïve—would also have to be considered completely pointless. Yet, this is what theorists that aim to avoid revenge via Field's argument ought to believe: as the foregoing discussion shows, they ought to believe that notions whose extension can be determined with a precise criterion *but trivialize* T are just not understandable enough.[21] I now highlight the consequences of C-EXTENSION with a concrete, and well-known, example:

Incompatible negations Let us consider again the logic K3 (see page 4). K3-logical validity is defined as preservation of value **1** in every partial evaluation. However, other logics can be defined from partial evaluations that are weak enough to support forms of naïveté: the *logic of paradox*, LP, is a case in point.[22] A sentence φ is a LP-consequence of a set of sentences Γ if, for every partial evaluation p, if $p(\Gamma) = \mathbf{1}$ or **1/2**, then $p(\varphi) = \mathbf{1}$ or **1/2**. K3- and LP-based theories are dual in several respects. In particular, K3-based theories do not satisfy all instances of the law of excluded middle (LEM) $\varphi \vee \neg\varphi$, but satisfy all instances of *ex falso quodlibet* (EFQ) $\varphi \wedge \neg\varphi \vdash \psi$, while LP-based theories validate all the instances of LEM, but fail to validate all the instances of EFQ. Since the K3-negation trivialize LP-based theories (and *vice versa*), accepting C-EXTENSION would force advocates of K3-based theories to declare the LP-negation unintelligible (and *vice versa*), even though both excluded middle and *ex falso quodlibet* are classically valid and customarily employed in (formalized) mathematical reasoning.

Can C-EXTENSION be defended even after one realizes its implications for the comparisons of theories of truth? The foregoing considerations should give the advocate of C-EXTENSION pause. However, there are far more conclusive reasons to

[21] Besides being conceptually problematic, the idea that comparing theories of truth is a futile exercise is disproven by the truth-theoretic literature. On the one hand, proponents of non-classical theories devote considerable efforts in exploring and highlighting the advantages of their approach over their rivals, classical and non-classical alike [see e.g. Beall (2009, Chapters 4–5), Field (2008, Chapters 7–8, 10–14, 24–26), Priest (2006, Chapter 20)]. On the other hand, advocates of classical theories of truth have sought to expose the limitations of non-classical approaches, from mathematical and from broadly methodological standpoints [see e.g. Halbach (2014, Chapter 20), Halbach and Nicolai (2017), Williamson (2017), Murzi and Rossi (2018a)].

[22] See Asenjo (1966) and Priest (1979).

reject this principle: in fact, C-EXTENSION seems to clash directly with the intelligibility of naïve truth. If C-EXTENSION is correct, it applies across the board, and not just to revenge-breeding semantic notions. Its application to naïve truth shows that this notion was not sufficiently understandable before the development of suitable non-classical theories in relatively recent times. At the very least, C-EXTENSION entails that naïve truth was not sufficiently understandable when it was first formulated as a limitative result (see e.g. Tarski 1936) in that it yields triviality once added to the accepted formal systems of first-order arithmetic (or some other sufficiently expressive theory) formulated in classical logic.

One might try to save C-EXTENSION postulating two senses of 'understanding'. In a 'shallow' sense, one understands semantic notions irrespectively of their extensions and of whether they trivialize other notions. In a 'deep' sense, understanding semantic notions requires C-EXTENSION to be satisfied. Some form of shallow/deep distinction is customary in scientific domains: non-specialists understand shallowly notions such as *computable*, *black hole*, or *DNA*, while specialists understand these notions properly. Understanding semantic notions properly might require to consult the relevant experts, while the shallow sense 'saves the appearances' according to which everyone can understand naïve truth and bivalent determinateness. However, in order to use C-EXTENSION to escape revenge paradoxes, one must suppose that only the deep sense is relevant to determine which semantic notions and paradoxes there are. But there are at least two problems with this supposition. First, the community of truth-theorists is split over which theory of truth is correct. Therefore, one accepting C-EXTENSION has to conclude that there are experts lacking a deep understanding of genuine semantic notions *just because they disagree* (for discussion, see Williamson 2007, Chapter 4). Second, this supposition is problematic for the reasons highlighted above. Before non-classical theories were available, naïve truth was only shallowly understandable but it generated genuine paradoxes, such as the Liar paradox. Why, then, should bivalent determinateness fail to generate genuine paradoxes?

One could further object that the understanding of naïve truth is and has always been (at least implicitly) coupled with some restriction of classical logic that avoids triviality. From this perspective, the understanding of naïve truth is and has always been perfectly in line with C-EXTENSION. Now, this objection might in effect provide a construal of C-EXTENSION that does not make it immediately implausible, although independent support would be needed to show that the understanding of naïve truth is indissolubly coupled with a suitable non-classical logic. However, this objection fails to vindicate Field's anti-revenge argument, for it is equally applicable to bivalent determinateness. If naïve truth trivializes classical theories but it is to be understood as an essentially non-classical concept, then the same could be said for bivalent determinateness: this notion trivializes several currently available non-classical theories but it is also to be understood as a (super-)non-classical concept that requires extremely weak non-classical theories. Therefore, this objection does not separate naïve truth from bivalent determinateness: it shows that, in both cases,

a non-classical logic that makes these notions non-trivial is required to claim an understanding of them. But this does not make bivalent determinateness any less acceptable than naïve truth.[23]

In conclusion, Field's argument does not offer an acceptable way to distinguish between naïve truth and bivalent determinateness so that the former can be claimed to generate genuine paradoxes while the latter can't. Since nothing in the foregoing arguments relies on the specificities of bivalent determinateness or Field's theory, I conclude that Field's argument does not offer a way out of revenge paradoxes motivated by the MT-semantics. In the next section, I consider a possible criterion to rule out MT-semantic notions, that implicitly informs several anti-revenge strategies, and argue that it also fails to provide a convincing base to distinguish 'revenge-breeding' from 'standard' paradoxical notions.

5 Model-theoretic instrumentalism

As Field's argument makes clear, naïve truth and D1–D5-bivalent determinateness are separated by their conceptual justification: the latter is motivated by the MT-semantics of an object-theory, while the former isn't. If one could show that the MT-semantics cannot motivate genuine semantic notions, one could argue that bivalent determinateness does not yield a genuine paradox. Some authors view the MT-semantics as a mere *instrument* to prove the object-theory non-trivial; as such, the MT-semantics might not allow one to carve out genuine semantic notions (see e.g. Beall (2007, pp. 10–11) and Yablo (2003, pp. 328–329)). Call this position MT-instrumentalism. One possible motivation for it is that different MT-semantics can be given in different meta-theories. Some MT-semantic notions definable in a classical meta-theory might not be definable in a *non-classical* meta-theory, including revenge-breeding notions such as the one employed in the paradox of Bivalent Determinateness.

Nevertheless, MT-instrumentalism does not provide a way out of the MT-revenge paradoxes. In short, even if the MT-semantics is regarded as a mere tool to prove the non-triviality of a theory T, there are crucial facts about the logical behaviour of T's sentences that can be captured in *any* MT-semantics for T. More precisely, *every* MT-semantics for T can distinguish sentences that respect all the principles of classical logic from sentences that obey distinctively non-classical principles. This shows that MT-semantics track genuine semantic distinctions, irrespectively of the specific models being employed, and thus irrespectively of the restrictions imposed on the extension of semantic notions from the set-theoretic nature of models.

[23] Alternatively, one might think that the classical logician has an understanding of naïve truth via her acceptance of a collection of instances of the naïve principles that does not yield triviality (see e.g. Horwich (1990), and McGee (1992), and Cieśliński (2007) for criticisms). However, this alternative objection would also fail to separate naïve truth from bivalent determinateness, thus failing to undermine the revenge problem posed by the latter notion.

I will again articulate this point in connection with Field's theory, although the argument generalizes easily. Field's theory F (sketched on page 4) is an infinite set of sentences that contains all the instances of several classical laws and of the T-SCHEMA (formulated with Field's conditional $\rightarrow_{\mathsf{F}}$), and it is closed under several classical rules of inference and INTER-SUBSTITUTIVITY. Recall, the set F is defined model-theoretically, namely via some evaluation function $v_{\mathscr{F}}$ from the sentences of the language to a set of semantic values. F consists of the sentences to which $v_{\mathscr{F}}$ assigns the value $\mathbf{1}$:

$$\mathsf{F} = \{\varphi \in \mathsf{SENT} \mid v_{\mathscr{F}}(\varphi) = \mathbf{1}\}$$

where SENT indicates the set of sentences of the language of Field's theory. Let's say that an evaluation function $v_{\mathscr{F}}$ from SENT to a set of semantic values *defines* F if $\mathsf{F} = \{\varphi \in \mathsf{SENT} \mid v_{\mathscr{F}}(\varphi) = \mathbf{1}\}$. For the MT-instrumentalist, *how* F is defined is completely insubstantial. In particular, F could be defined by an evaluation $v^n_{\mathscr{F}}$ defined in a non-classical meta-theory, in which it is not always the case that $v^n_{\mathscr{F}}(\varphi) = \mathbf{1}$ or $v^n_{\mathscr{F}}(\varphi) \neq \mathbf{1}$.

Now, F is a non-classical theory, and therefore the principles of classical logic hold for some of F's sentences, while other sentences exhibit a non-classical behaviour. Suppose $v^*_{\mathscr{F}}$ is an evaluation that defines F. In $v^*_{\mathscr{F}}$, we have

$$v^*_{\mathscr{F}}\big(\neg[\forall x(x = x) \leftrightarrow_{\mathsf{F}} \neg\forall x(x = x)]\big) = \mathbf{1} \quad \text{and} \quad v^*_{\mathscr{F}}\big(\lambda \leftrightarrow_{\mathsf{F}} \neg\lambda\big) = \mathbf{1}.$$

The biconditional of Field's logic $\leftrightarrow_{\mathsf{F}}$ is non-classical, so one should expect it to behave non-classically. Still, we know that $(\forall x(x = x) \vee \neg\forall x(x = x))$ is in F, and that if $\varphi \vee \neg\varphi$ is in F, then φ is closed under all the classical principles involving φ, and also the classical principles for the (otherwise non-classical) $\leftrightarrow_{\mathsf{F}}$. This is enough to conclude that, in any evaluation that defines F, the laws of identity and Liar sentences have different logical behaviours: while $\forall x(x = x)$ obeys all the classical principles (including $\neg(\varphi \leftrightarrow_{\mathsf{F}} \neg\varphi)$) because it obeys the enabling condition by which $\leftrightarrow_{\mathsf{F}}$ behaves just like the classical biconditional on it, λ obeys the non-classical principle $\varphi \leftrightarrow_{\mathsf{F}} \neg\varphi$.[24,25]

Another way to discern the non-classical logical behaviour of some sentences in F uses Field's very notion of D1-D4-determinateness: in *any* evaluation function $v^*_{\mathscr{F}}$ that defines F, $v^*_{\mathscr{F}}\big(\mathsf{Det}(\forall x(x = x))\big) = \mathbf{1}$ and $v^*_{\mathscr{F}}\big(\neg\mathsf{Det}(\lambda)\big) = \mathbf{1}$. A moment's reflection shows that some grasp of the non-classicality of sentences such as λ is at least presupposed for Field's theory. For, it if weren't, what could have motivated a determinateness operator that describe its 'indeterminateness' in the first place?[26]

[24] In Murzi and Rossi (2018a), purely object-linguistic revenge paradoxes are developed that exploit the characterization of sentences obeying, or failing to obey, all the principles of classical logic.

[25] Were the meta-theory in which $v^*_{\mathscr{F}}$ is constructed sufficiently strong to prove $v^*_{\mathscr{F}}(\lambda \vee \neg\lambda) \neq \mathbf{1}$, the non-classical behaviour of λ would be more evident. However, that $v^*_{\mathscr{F}}(\lambda \vee \neg\lambda) \neq \mathbf{1}$ is not a fact about F, but about what is *not* in F, and it might be out of reach for non-classical meta-theories. Nevertheless, the facts mentioned above suffice to conclude that λ and $\forall x(x = x)$ have different logical behaviours.

[26] Accepting some form of non-classicality for λ seems to be a prerequisite to determine the attitudes a rational agent should have towards λ (see Caie 2012).

The foregoing observations show that any MT-semantics tracks crucial object-theoretic facts. This does not depend on whether one attributes an instrumental role to the MT-semantics, and it is also independent on the specific evaluation functions employed, and thus on the specific models on which they are defined. Therefore, even the MT-instrumentalist has to concede that the MT-semantics, whether developed in a classical or a non-classical meta-theory, can be used to articulate genuine semantic distinctions. This lesson finds a natural application to revenge paradoxes: they make explicit, via paradoxical notions, some semantic distinctions that can be made, if only partially, in the MT-semantics. For instance, bivalent determinateness makes explicit the different logical behaviour of sentences such as λ and $\forall x(x = x)$.

An advocate of Field's theory (or of another sufficiently expressive theory) might object at this point that the different logical behaviours of λ and $\forall x(x = x)$ can already be captured by a determinateness operator obeying D1-D4 alone, since both $\mathsf{Det}(\forall x(x = x))$ and $\neg\mathsf{Det}(\lambda)$ are in F. If one then tries to formulate revenge-paradoxical sentences, the objection continues, their non-classical behaviour can be captured iterating D1-D4-determinateness. Recall the sentence λ^* equivalent to $\neg\mathsf{True}(\ulcorner\mathsf{Det}(\lambda^*)\urcorner)$, mentioned on page 5: its non-classical behaviour cannot be captured by declaring it 'not determinately true', since it is equivalent to λ^*. However, Field's theory declares λ^* to be 'not determinately determinately true', since $\neg\mathsf{Det}(\mathsf{Det}(\mathsf{True}(\ulcorner\lambda\urcorner^*)))$ is in F. And so on. However, Welch (2014) has shown that if the iteration is continued long enough, 'ineffable Liars' can be defined in F, namely sentences whose non-classical behaviour cannot be captured by *any* iteration of determinateness that can be constructed in F (see also Rayo and Welch 2007). Yet, the non-classical behaviour of Welch's ineffable Liars can be characterized very easily in several MT-semantics for F, simply using the notion of bivalent determinateness.

In conclusion, the MT-semantics might be thought to have primarily an instrumental role, but it also provides crucial information about the behaviour of sentences of the object-theory. For this reason, the MT-semantics can be used to carve out genuine semantic notions, that make such behaviour explicit. These notions may be paradoxical, but cannot be thought of as the by-product of a mere instrument.[27]

[27] The MT-instrumentalist could object that the very idea of *defining* an object-theory via evaluation functions, i.e. MT-semantic notions, is misguided. For instance, Field (2008, p. 277) argues that the MT-semantics could be seen just as offering 'a proof that one won't get into trouble operating with the inferences in it', although what is validated by the MT-semantics 'outruns 'real validity''. However, such a position offers no way out of the foregoing arguments for the significance of MT-notions. Even if one could select a subset F_0 of F which is simple enough to be characterized axiomatically, the distinctions concerning the logical behaviour of F's sentences that can be made in the MT-semantics would apply to F_0's sentences just as well.

6 Concluding remarks

Piet Hein famously quipped that 'problems worthy of attack prove their worth by hitting back'. The growing attention which truth theorists devote to revenge paradoxes witnesses that they are considered to be worthy of attack. I hope to have shown that they do hit back.

In particular, Field's anti-revenge-strategy does not offer a way out of the MT-revenge paradoxes. The relevance of Field's strategy is hardly over-estimated, since it is applicable to every theory of truth that has an MT-semantics, and it threatens to undermine the coherence of MT-semantic notions, or notions motivated by the MT-semantics, more generally. However, Field's defence, once fully articulated, offers no acceptable reason to think that revenge-breeding notions are not genuine semantic notions. MT-instrumentalism is equally unsuccessful as a strategy to reject revenge-breeding MT-notions, for they can be seen to carve out genuine semantic distinctions about the object-theories.

As emphasized in Sect. 1, the foregoing arguments do not show that every MT-revenge paradox is successful; they show that MT-revenge paradoxes are not to be dismissed *just because* they involve the MT-semantics. However, this modest conclusion has implications for our understanding of semantic paradoxes, and ultimately for deciding amongst approaches to treating truth and other semantic notions in the object-language. One such implication is that there is no well-motivated ground to distinguish between 'standard' and 'revenge' paradoxes. Cogently motivated revenge-breeding notions are legitimate semantic notions, just like naïve truth. Revenge paradoxes are not *ex post facto* antinomies, that only arise 'after' a solution to the 'standard' paradoxes has been found, and target that specific solution: they are expressive limitations that have been affecting the language all along, exactly as the 'standard' paradoxes. The whole distinction between 'revenge' and 'standard' paradoxes should be abandoned.

But if there is no such a thing as a 'standard' versus 'revenge' distinction, then a theory of truth that treats successfully the 'standard' paradoxes but fails to treat the 'revenge' ones is just a theory that solves the paradoxes only partially. For example, if naïve truth and bivalent determinateness are equally legitimate notions, the very logical revision that is operated to express the former is not fully justified if it cannot also be used to express the latter. So, abandoning the 'standard' versus 'revenge' distinction can help decide among theories of truth and treatments of paradoxes.[28]

One might worry that rehabilitating revenge paradoxes would have devastating consequences. Since revenge paradoxes are so pervasive, is there still hope for theories that express all the genuine semantic notions? Field himself seems to share such a worry:

> The strong revenge worry is that adding such an operator [bivalent determinateness] to the language would produce a new paradox that requires

[28] One could object that the revenge-breeding notions are somewhat less important than truth (see e.g. Maudlin 2007). Yet, the discussion in Sect. 5 makes it clear that revenge-breeding notions can express key facts of the object-theories.

> giving up the truth schema. Substantiating this would be a fatal blow to any claim that a G-solution [the kind of solutions he favours] adequately resolves all the paradoxes. (Field 2007, p. 120)

However, there is no a priori reason to think that no theory can express all the semantic notions in the object-language. The semantic paradoxes—'standard' and 'revenge' alike—simply show that deep expressive limitations stand on the way to this goal. What cannot be reasonably hoped for is a theory of truth and other semantic notions that is completely free from expressive limitations. What can be reasonably hoped for is a theory that addresses such expressive limitations *uniformly*—thus applying the same solution to 'standard' and 'revenge' paradoxes. From this point of view, some classical theories seem promising: in particular, *contextualist theories* can be given that offer the very same solution to *all* semantic paradoxes, while retaining the strength of full classical logic.[29]

Acknowledgements Open access funding provided by Paris Lodron University of Salzburg. I am grateful to Volker Halbach and Hartry Field for helpful exchanges on revenge paradoxes and model-theoretic semantics, and to an anonymous referee for useful comments. Special thanks are due to Julien Murzi for detailed feedback and extensive discussions that have led to significant improvements.

FundingI am grateful to the Austrian Science Fund (FWF), Grant No. P2971-G24, for generous financial support during the time this paper was written.

References

Asenjo, F. G. (1966). A calculus of antinomies. *Notre Dame Journal of Formal Logic*, *7*(1), 103–105.
Bacon, A. (2013). Non-classical metatheory for non-classical logics. *Journal of Philosophical Logic*, *42*(2), 335–355.
Bacon, A. (2015). Can the classical logician avoid the revenge paradoxes? *Philosophical Review*, *124*(3), 299–352.
Beall, J. (2006). True, false and paranormal. *Analysis*, *66*(2), 102–114.
Beall, J. (2007a). Prolegomenon to future revenge. In J. Beall (Ed.), *Revenge of the Liar* (pp. 1–30). Oxford: Oxford University Press.
Beall, J. (Ed.). (2007b). *Revenge of the Liar*. Oxford: Oxford University Press.
Beall, J. (2009). *Spandrels of truth*. Oxford: Oxford University Press.
Blamey, S. (2002). Partial logic. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (2nd ed., Vol. V, pp. 261–353). Dordrecht: Kluwer Academic Publishers.
Caie, M. (2012). Belief and indeterminacy. *Philosophical Review*, *121*(1), 1–54.
Chemla, E., Égré, P., & Spector, B. (2017). Characterizing logical consequence in many-valued logic. *Journal of Logic and Computation*, *27*(7), 2193–2226.
Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: Introduction to semantics* (2nd ed.). Cambridge: MIT Press.

[29] See Glanzberg (2004), Parsons (1974), Simmons (1993, 2015) and Murzi and Rossi (2018b).

Cieśliński, C. (2007). Deflationism, conservativeness and maximality. *Journal of Philosophical Logic*, *36*, 695–705.
Cook, R. (2007). Embracing revenge: on the indefinite extensibility of language. In J. Beall (Ed.), *Revenge of the Liar* (pp. 31–52). Oxford: Oxford University Press.
Davidson, D. (1967). Truth and meaning. *Synthese*, *17*, 304–323.
Dummett, M. (1993). *Frege: Philosophy of language* (Second ed.). Cambridge, MA: Harvard University Press.
Eklund, M. (2007). The liar paradox, expressibility, possible languages. *The revenge of the Liar* (pp. 53–77). Oxford: Oxford University Press.
Field, H. (1994). Deflationist views of meaning and content. *Mind*, *103*, 249–285.
Field, H. (2002). Saving the truth schema from paradox. *Journal of Philosophical Logic*, *31*(1), 1–27.
Field, H. (2003). A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, *32*, 139–177.
Field, H. (2007). Solving the paradoxes, escaping revenge. In J. Beall (Ed.), *Revenge of the Liar* (pp. 53–144). Oxford: Oxford University Press.
Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.
Field, H. (2013). Naive truth and restricted quantification: Saving truth a whole lot better. *Review of Symbolic Logic*, *7*(1), 147–191.
Field, H., Lederman, H., & Øgaard, T. (2017). Prospects for a naïve theory of classes. *Notre Dame Journal of Formal Logic*, *58*(4), 461–506.
Fischer, M., Halbach, V., Kriener, J., & Stern, J. (2015). Axiomatizing semantic theories of truth? *The Review of Symbolic Logic*, *8*, 257–278.
Glanzberg, M. (2004). A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic*, *33*, 27–88.
Glanzberg, M. (2014). Explanation and partiality in semantic theory. In A. Burgess & B. Sherman (Eds.), *Metasemantics: New essays on the foundations of meaning*. Oxford: Oxford University Press.
Halbach, V. (2014). *Axiomatic theories of truth* (2nd ed.). Cambridge: Cambridge University Press.
Halbach, V., & Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic*, *71*, 677–712.
Halbach, V., & Nicolai, C., (2017). On the costs of nonclassical logic. *Journal of Philosophical Logic*. https://doi.org/10.1007/s10992-017-9424-3.
Hill, C. (2014). *Meaning, mind, and knowledge*. Oxford: Oxford University Press.
Horsten, L. (2009). Levity. *Mind*, *118*(471), 555–581.
Horwich, P. (1990). *Truth*. Oxford: Oxford University Press.
Jech, T. (2002). *Set theory, the third* (Millennium ed.). Berlin: Springer.
Ketland, J. (2003). Can a many-valued language functionally represent its own semantics? *Analysis*, *63*(4), 292–297.
Kleene, S. C. (1952). *Introduction to metamathematics*. Amsterdam: North-Holland.
Kremer, M. (1988). Kripke and the logic of truth. *Journal of Philosophical Logic*, *17*(3), 327–332.
Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, *72*, 690–716.
Leitgeb, H. (2007). On the metatheory of Field's 'Solving the paradoxes, escaping revenge'. In J. Beall (Ed.), *Revenge of the Liar* (pp. 159–183). Oxford: Oxford University Press.
Lepore, E. (1983). What model-theoretic semantics cannot do. *Synthese*, *54*(2), 167–187.
Maudlin, T. (2007). Reducing revenge to discomfort. In J. Beall (Ed.), *Revenge of the Liar* (pp. 184–196). Oxford: Oxford University Press.
McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic*, *21*, 235–241.
McGee, V. (2016). Thought, thoughts, and deflationism. *Philosophical Studies*, *173*, 3153–3168.
Murzi, J., & Rossi, L. (2018a). Generalized revenge **(under review)**.
Murzi, J., & Rossi, L. (2018b). Reflection principles and the Liar in context. forthcoming in Philosophers' Imprint
Parsons, C. (1974). The Liar paradox. *Journal of Philosophical Logic*, *3*(4), 381–412.
Priest, G. (1979). The logic of paradox. *Journal of Philosophical Logic*, *8*, 219–241.
Priest, G. (2006). *Contradiction* (expanded edition). Oxford: Oxford University Press.
Priest, G. (2007). Revenge, Field, and ZF. In J. Beall (Ed.), *Revenge of the Liar* (pp. 225–233). Oxford: Oxford University Press.
Rayo, A., & Welch, P. (2007). Field on revenge. In J. Beall (Ed.), *Revenge of the Liar* (pp. 234–249). Oxford: Oxford University Press.

Restall, G. (2007). Curry's revenge: The costs of non-classical solutions to the paradoxes of self-reference. In J. Beall (Ed.), *Revenge of the Liar* (pp. 262–271). Oxford: Oxford University Press.
Rosenblatt, L. (2015). Two-valued logics for transparent truth theory. *Australasian Journal of Logic*, *12*(1), 44–66.
Rossi, L. (2016). Adding a conditional to Kripke's theory of truth. *Journal of Philosophical Logic*, *45*(5), 485–529.
Scharp, K. (2007). Aletheic vengeance. In J. Beall (Ed.), *Revenge of the Liar* (pp. 272–319). Oxford: Oxford University Press.
Scharp, K. (2013). *Replacing truth*. Oxford: Oxford University Press.
Shapiro, L. (2011). Expressibility and the liar's revenge. *Australasian Journal of Philosophy*, *89*(2), 1–18.
Simmons, K. (1993). *Universality and the Liar: An essay on truth and the diagonal argument*. Cambridge: Cambridge University Press.
Simmons, K. (2007). Revenge and context. In J. Beall (Ed.), *Revenge of the Liar* (pp. 345–367). Oxford: Oxford University Press.
Simmons, K. (2015). Paradox, repetition, revenge. *Topoi*, *34*(1), 121–131.
Tarski, A. (1936). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica Commentarii Societatis Philosophicae Polonorum, 1*, 261–405, reprinted as 'The Concept of Truth in Formalized Languages' in Tarski A (1956), pp. 152–278.
Tarski, A. (1956). *Logic, semantics, metamathematics*. Oxford: Oxford University Press.
Welch, P. (2014). Some observations on truth hierarchies. *The Review of Symbolic Logic*, *7*(1), 1–30.
Williamson, T. (2007). *The Philosophy of Philosophy*. Oxford: Wiley-Blackwell.
Williamson, T. (2017). Semantic paradoxes and abductive methodology. In: B. Armour-Garb (Ed.), *The relevance of the Liar* (pp. 325–346). Oxford: Oxford University Press.
Yablo, S. (2003). New grounds for naive truth theory. In J. Beall (Ed.), *Liars and heaps. New essays on paradox* (pp. 312–330). Oxford: Oxford University Press.
Zardini, E. (2011). Truth without contra(di)ction. *Review of Symbolic Logic*, *4*(4), 498–535.