Journal of Philosophical Logic (2019) 48: 155–188
https://doi.org/10.1007/s10992-018-9483-0

Instability and Contraction

Méditations hégéliennes I

Elia Zardini[1,2]

Received: 28 March 2017 / Accepted: 19 February 2018 / Published online: 3 January 2019

Abstract
In other works, I've proposed a solution to the semantic paradoxes which, at the technical level, basically relies on failure of contraction. I've also suggested that, at the philosophical level, contraction fails because of the instability of certain states of affairs. In this paper, I try to make good on that suggestion.

Keywords Causation · Contraction · Implication · Instability · Logical consequence · Revision theory · Semantic paradoxes · States of affairs

> An argument is a discourse where, having laid down certain conditions, something different from the underlying conditions necessarily obtains the conditions being what they are. And by 'the conditions being what they are' I mean that the consequence obtains because of the conditions, where by 'obtains because of the conditions' I mean in turn that there is no need of any further term for the coming into being of the necessity. (Aristotle, *Prior Analytics*, A, 1, 24b, my translation)

1 Contraction

Perhaps surprisingly, the *semantic paradoxes* can be *technically blocked* basically by giving up the principle of *contraction* (e.g. [42]), according to which, for every sentence φ and ψ, the fact that φ, φ (*i.e.* φ taken twice as a premise) entails ψ implies

✉ Elia Zardini
elia.zardini@campus.ul.pt

[1] LanCog, Language, Mind and Cognition Research Group, Philosophy Centre, University of Lisbon, Lisbon, Portugal

[2] International Laboratory for Logic, Linguistics and Formal Philosophy, School of Philosophy, National Research University Higher School of Economics, Moscow, Russian Federation

the fact that φ entails ψ (contraction *in the premises*) and the fact that φ entails ψ, ψ (*i.e.* ψ taken twice as a conclusion) implies the fact that φ entails ψ (contraction *in the conclusions*).[1] But how can one make *philosophical sense* of the failure of such principle?[2] Picking up on a suggestion I've recently put forth, and making use

[1]On reflection, this should not be that surprising. Typical semantic paradoxes rely on (variations of) the fact that, by selfreference and the properties of truth (*cf* [51], pp. 574–575), a sentence π is *double-faced* in that it is equivalent with $\ldots\ \pi\ \ldots$, where, for every φ, φ and $\ldots\ \varphi\ \ldots$ *together* entail some unwarranted consequence (*cf* fn 64). Typical semantic paradoxes then exploit *both of* π*'s faces* to infer that π *itself* entails the relevant unwarranted consequence, from which entailment $\ldots\ \pi\ \ldots$ can in turn be inferred *etc.* However, if contraction fails, the former inference is invalid and the semantic paradox is thereby blocked.

[2]Under widely shared assumptions, failure of contraction for φ is equivalent with φ & φ*'s being stronger than* φ. Not unusually, amused puzzlement is expressed as to how φ & φ can ever be stronger than φ. But such amusement is misplaced: natural language offers a wealth of cases where φ & φ would indeed seem stronger than φ: 'I have 1 EUR and I have 1 EUR' would seem to entail 'I have 2 EUR' and so be stronger than 'I have 1 EUR', 'It was raining and it was raining' would seem to entail 'It was raining for a while' and so be stronger than 'It was raining', 'I love you and I love you' would seem to entail 'I love you a lot' and so be stronger than 'I love you'... Therefore, it should be no big mystery *that* contraction fails (also in cases that have little to do with the semantic paradoxes)—the task of this paper is to provide a hopefully illuminating explanation of *why* it fails (at least in the case of the semantic paradoxes). Thanks to Pilar Terrés for discussions that brought about this fn.

[3]To the best of my knowledge, something along the lines of this thought has first been adumbrated by [14], who also would seem to relate *implication with causation* and *failure of contraction with instability*. However, he embeds these insights into a *resource*-theoretic framework that I regard as problematic. A typical example in this connection is the idea that, assuming that 1 cigar costs 2 EUR, while '1 EUR', '1 EUR' entails '1 cigar', '1 EUR' does not. As I understand it, this is explained by saying that, in '1 EUR', '1 EUR', the first occurrence of '1 EUR' represents a token of 1 EUR and the second occurrence of '1 EUR' represents another token of 1 EUR, so that, summing up, '1 EUR', '1 EUR' represents a token of 2 EUR, which, by assumption, does suffice for a token of 1 cigar; on the other hand, '1 EUR' represents a token of 1 EUR, which, by assumption, does not suffice for a token of 1 cigar ([14], p. 2). Such explanation is *in itself* problematic, as it unwarrantedly assumes that the token of 1 EUR represented by the second occurrence of '1 EUR' is *distinct* from the token of 1 EUR represented by the first occurrence of '1 EUR' (notice that it cannot be assumed that *new occurrences always represent new tokens*, for otherwise '1 EUR', 'Not 1 EUR' would no longer be inconsistent and '1 EUR' would no longer entail '1 EUR'). The explanation is also problematic in that, now granting for the sake of argument that it does work for linguistic expressions (such as '1 EUR') representing *resources*, it cannot easily be *extended* to linguistic expressions (such as 'Snow is white') representing *states of affairs* (*i.e.* to *declarative sentences*). What, for example, would 'Snow is white', 'Snow is white' represent? Presumably, two tokens of the state of affairs that snow is white, but how can these amount to *anything more than the simple fact* that snow is white? The problem is then that also one token of the state of affairs that snow is white, which is what 'Snow is white' would represent, amounts to the simple fact that snow is white, so that no relevant intelligible distinction between 'Snow is white', 'Snow is white' and 'Snow is white' could be drawn. A suggestion in the direction of such extension is made by [27] (and goes back at least as far as [29], p. 26), who propose that 'Snow is white', 'Snow is white' represents not two tokens of the *state of affairs* that snow is white, but two tokens of the *information* that snow is white (where, I add, information is in turn so understood that two tokens of the same information must have different *sources*). Such information-theoretic approach to failure of contraction comes fraught with the same problem about the distinctness assumption which afflicts the resource-theoretic approach. And, now granting for the sake of argument that assumption, while the information-theoretic explanation at least does manage to draw an intelligible distinction between 'Snow is white', 'Snow is white' and 'Snow is white' which is relevant for reasoning, it would seem both *too narrow* and *too wide* a distinction: too narrow because it is only relevant for *nondeductive* rather than *deductive* reasoning (since the distinction is only effective for nonconclusive sources, but then, for every natural number i, nothing relevant can be deductively inferred from the fact that, according to i nonconclusive sources, snow is white), and too wide because it applies to *just about every* piece of information (so that, if the information-theoretic explanation managed to justify failure of

of ideas long developed in the tradition of revision theory, this paper explores the thought that *contraction fails because of the instability of certain states of affairs*.[3,4]

contraction for a problematic sentence like the Liar sentence, it would also justify failure of contraction for an unproblematic sentence like 'Snow is white'). Notice also that both the resource-theoretic and the information-theoretic approach to failure of contraction would seem to apply *directly* only to failure of contraction in the *premises* rather than also to failure of contraction in the *conclusions*. Having said all this by way of providing some *contrastive background* for this paper's own attempt, I hasten to add that the above considerations are simply offered in the spirit of stimulating *further investigations into the philosophical foundations* of the resource-theoretic and information-theoretic approaches to failure of contraction: I'd hope that, just as with virtually all other logical principles, *there is more than one way of justifying failure of contraction*, and that those two approaches do point to two other such ways. Thanks to Aurélien Darbellay, Dan López de Sa and Sven Rosenkranz for suggestions concerning the material in this fn.

[4]In addition to the approaches to failure of contraction discussed in fn 3, a recent alternative approach (which has the distinctive feature of trying to preserve the assumption that premises and conclusions are put together into *sets*) is offered by [37], whose basic idea is that, if φ is *different from itself qua* having contradictory properties F and G, $\{\varphi, \varphi\}$ has an F member and a different G member. However, *for exactly the same reason*, $\{\varphi\}$ too has an F member and a different G member, so that no difference between $\{\varphi, \varphi\}$ and $\{\varphi\}$ has yet been made out. Moreover, the *reason* Weber alleges for why, in the relevant cases, φ is different from itself threatens the stability of his overall position. To wit, Weber claims that, in the relevant cases, $\varphi, \varphi \vdash \psi$ holds only because $\varphi^i, \varphi^j \vdash \psi$ holds (where, in the relevant *hierarchy*, φ^i is of rank i and φ^j of rank j), so that, while we can "drop the indices" and maintain that $\varphi, \varphi \vdash \psi$ holds (somehow glossing over the fact that, while in the relevant cases every premise in φ, φ is selfreferential, one premise in φ^i, φ^j is not), such drop has the effect of making φ different from itself (presumably, because it originates from φ^i and it originates from φ^j, and everything originating from φ^i is different from everything originating from φ^j). The problem is that such explanation, if good, applies equally well to the case where $\varphi \vdash \psi, \psi$ holds only because $\varphi \vdash \psi^i, \psi^j$ holds, thereby predicting that $\varphi \vdash \psi$ does not hold, whereas Weber's overall position crucially relies on a principle (the metarule of *reasoning by cases*, amply touched on in Section 5) which is virtually incompatible with such prediction (acceptance of contraction in the conclusions is another distinctive feature of Weber's approach to failure of contraction). Similarly, an analogous explanation relying on a hierarchy of *negations*—rather than on a hierarchy of *implications* as Weber's does—would predict that contraction fails also for a *Liar sentence*, whereas Weber's overall position crucially relies on principles (the law of *excluded middle* and the metarule of *reasoning by cases*) which are virtually incompatible with such prediction (indeed, with failure of contraction *of any sentence intersubstitutable with its negation*: if φ is any such sentence, by excluded middle $\varphi \vee \neg\varphi$ is a logical truth, and so, since φ is intersubstitutable with $\neg\varphi$, $\varphi \vee \varphi$ is a logical truth, and hence, by reasoning by cases, φ is a logical truth; for good measure, since, by excluded middle, $(\varphi \vee \neg\varphi)$ & $(\varphi \vee \neg\varphi)$ is a logical truth just as well, so is φ & φ, with fn 54 indicating that the real problem here is reasoning by cases rather than excluded middle). Another recent alternative approach to failure of contraction is offered by [33], who interprets $\varphi_0, \varphi_1, \varphi_2 \ldots, \varphi_i \vdash \psi$ as being tantamount to $\varphi_0 \Rightarrow (\varphi_1 \Rightarrow (\varphi_2 \ldots \Rightarrow (\varphi_i \Rightarrow \psi)))\ldots)$ (where $\Rightarrow$ is *logical implication*), for then suggesting that one can resist the transition from $\varphi \Rightarrow (\varphi \Rightarrow \psi)$ to $\varphi \Rightarrow \psi$. However, a little reflection on $\varphi \Rightarrow (\varphi \Rightarrow \psi)$ shows that the transition is very hard to resist: if φ logically implies $\varphi \Rightarrow \psi$, since it also logically implies φ it seems that it logically implies both premises of a $\Rightarrow$-based *modus-ponens* argument with ψ as conclusion; if so, by transitivity of logical implication, φ would logically imply ψ. To anticipate, on the view to be explored in this paper, the transition is resisted because, while φ logically implies *either* premise, it does not logically imply *both* (Section 5). Presumably, and *pace* [45], p. 584, on Shapiro's overall position, φ does logically imply both premises (as it logically implies their conjunction), so that the transition is even harder to resist. Shapiro resists it by in effect denying *modus ponens as traditionally understood* (*i.e.* as concerning the validity of the argument from both an implication and its antecedent to its consequent). Notice that Shapiro does accept that "$\varphi \Rightarrow \psi, \varphi \vdash \psi$" holds, but that's simply because, on his interpretation of premise structure, that claim is tantamount to $(\varphi \Rightarrow \psi) \Rightarrow (\varphi \Rightarrow \psi)$, which, while undeniably true, has little to do with the traditional idea of *modus ponens*. In fact, Shapiro's interpretation of premise structure would seem to fail to capture the idea that an argument concerns *the assumption of no more and no less than all its premises together* (see [45], pp. 581–582 for an indication of the relevant facts and [53] for some subtleties about such assumption), since,

A bit more in detail, such exploration will proceed in five main steps. In Section 2, I'll lay out the salient properties and structure of the domain of states of affairs. In Section 3, I'll characterise the nature, and support the existence, of unstable states of affairs. In Section 4, I'll develop a theory of truth-related causation where instability is compatible with contraction. In Section 5, I'll show how to extract a theory of truth-related implication from that theory of causation and indicate how features that are distinctive of implication *vis-à-vis* causation make instability incompatible with contraction. In Section 6, I'll show how to turn that theory of implication into a semantics with respect to which a natural noncontractive logic (spoiler in the appended fn)[5] which blocks the semantic paradoxes is sound and complete. All in all, such exploration will thus vindicate the thought that *contraction fails when, out of an underlying relation of causation between states of affairs in whose context contraction is still compatible with the instability of some states of affairs, we extract a relation of implication in whose context contraction becomes tantamount to the denial that any state of affairs is unstable*.

2 States of Affairs

Since the thought to be explored grounds facts about *implication* (and so about *logical consequence*) in facts about *causation*, it makes sense to focus on a domain of objects that can stand *both* in causal *and* in implicational relations, and *states of affairs* (SAs) are a natural choice in this respect.[6] As usual (e.g. [38]), I assume SAs to be *relatively fine-grainedly individuated* objects that *exist relatively independently of how*

on his interpretation, $\varphi, \varphi \Rightarrow \psi \vdash \psi$—contrary to the commuted $\varphi \Rightarrow \psi, \varphi \vdash \psi$—does not hold (thereby showing that an argument concerns the assumption of also an *order* of its premises, in violation of the "no-more"-component of the idea), nor does $\varphi, \psi \vdash \varphi \,\&\, \psi$ (thereby showing that an argument does not concern the assumption of *all its premises together*, in violation of the "no-less"-component of the idea). But it is *precisely for that idea of argument*—arguably, the one that is after all *most directly involved in informal presentations of the semantic paradoxes*—that this paper wants to make philosophical sense of failure of contraction. Shapiro does also consider an interpretation of premise structure which captures the idea that an argument concerns the assumption of no more and no less than all its premises together (by letting $\varphi_0, \varphi_1, \varphi_2 \ldots, \varphi_i \vdash \psi$ be tantamount to $(\varphi_0 \,\&\, \varphi_1 \,\&\, \varphi_2 \,\ldots\, \&\, \varphi_i) \Rightarrow \psi$), for then doubting that one can resist the transition from $(\varphi \,\&\, \varphi) \Rightarrow \psi$ to $\varphi \Rightarrow \psi$. However, even before recourse to heavy-duty explanations like the one offered in this paper, *that* transition is indeed easily resisted (fn 2). While a remark analogous to the one in fn 3 applies concerning the spirit in which the above considerations about Weber's and Shapiro's approaches to failure of contraction are offered, a general point emerging from these regards the difficulty of *motivating* a noncontractive approach to Curry's paradox which does not extend to the Liar paradox (in addition to the *intrinsic* problems facing any such differential approach, see [48]). Thanks to an anonymous referee for encouraging a development of the material in this fn.

[5]The $\Rightarrow$-free fragment of [46]'s $\mathbf{IKT}_{\Rightarrow\mathrm{tf}}$.

[6]Going back at least as far as Aristotle, *Categories*, 10, 12b, the notion of a SA has a rich history (e.g. [34]), and it is interesting to observe that such history already includes an attempt [31] at grounding facts about logical consequence in facts about SAs. Thanks to John Horden and an anonymous referee for pushing me to be more explicit about this background.

the world is.[7] To wit, as for the first property, I want to allow for the SA ↑Socrates is a better philosopher than Cicero↑[8] to be the same as ↑Socrates is a better philosopher than Tully↑, whereas I don't want to allow for it to be the same as ↑Socrates is a better philosopher than Cicero and tigers are animals↑; as for the second property, I want to allow for the existence of ↑Socrates is wise↑ to depend on Socrates' existing, whereas I don't want to allow for its existence to depend on Socrates' being wise (I'll briefly mention how these issues reflect on logical consequence in fn 63). What does systematically depend on how the world is whether a SA *obtains* or not. For example, ↑Socrates is wise↑ obtains iff Socrates is wise. *Facts* are obtaining SAs: there exists the fact that P iff $\uparrow P\uparrow$ obtains.

I assume that sentences can be interpreted as *expressing SAs*. Given this semantic work that SAs are supposed to do, and given that sentences include a class of atomic sentences and are closed under the operators of negation, conjunction and disjunction, I correspondingly assume that SAs include a class of *simple* SAs and are closed under the involutive 1ary operation of *complementation* (so that, for every SA s, s^* is the complementation of s) and the associative and commutative 2ary *composition* operations of *combination* and *alternation* (so that $s\hat{}t$—st for short—and $s|t$ are, respectively, the combination and alternation of s and t).[9] The obtaining of s^* is supposed to consist in s's in some way or other *failing* to obtain, the obtaining of st is supposed to consist in the obtaining *together as features* of s and t (in other words, in the *coobtaining* of s and t) and the obtaining of $s|t$ is supposed to consist in the obtaining *together as possibilities* of s and t (in other words, in the *contraobtaining* of s and t), which I take to capture important notions of negation, conjunction and disjunction respectively (see in particular [47] for some discussion about negation and [45] for some discussion about conjunction). I also assume the existence of two 0ary operations: the *absolutely positive* SA (**p**) and the *absolutely negative* SA (**n**), whose further explanation (as well as the introduction of a further, *truth-theoretic* operation) is better postponed until Section 4.[10]

[7] I'll assume a lot of other things as well (about SAs, instability, causation, implication and logical consequence) which, while not particularly implausible, ideally should be supported by some argument. Still, my aim in this paper is *not* to *demonstrate from first principles that contraction fails*, but only to *trace an attractive* (to me at least!) *route through a complex terrain which would make sense of such failure*. Given this aim, I focus in the paper on *opening up the route* rather than on *buttressing its background*, and consequently leave unargued many assumptions of the operative framework. Their defence will have to wait for another day; in the meanwhile, the fact that they are not particularly implausible will hopefully suffice not to detract excessively from the interest of what the paper does offer. Thanks to an anonymous referee for urging me to be clear about this.

[8] Throughout, I use $\uparrow P\uparrow$ to denote the SA of its being the case that P.

[9] Throughout, I assume that * binds more strongly than $\hat{}$ and this binds more strongly than $|$, so that e.g. $s|tu^*$ is to be read as $s|(t(u^*))$.

[10] Throughout, I assume that *complex SAs are fine-grainedly individuated*, so that different operations or different SAs as arguments yield different SAs as values (insofar as that is compatible with involutivity of complementation as well as with associativity and commutativity of the composition operations—*plus*, only to the limited extent explained at the end of the development in Section 4, their idempotency).

3 Instability

A prominent relation SAs can enter into is *causation*: some SAs cause some SAs, in the sense that the *obtaining* of some SAs causes the *obtaining* of some SAs.[11] I follow the *pre-Humean* tradition in employing a *broad* notion of causation, which applies to *every* case where some facts *determine* some facts (e.g. Plato, *Phaedo*, 96a–102a) rather than only to cases of *natural efficient* causation (see [32] for a recent survey of the historical emergence of the latter notion).[12] Thus, to give some examples of different kinds, just as the SA w ↑The wood is burning↑ causes the SA a ↑Only ashes remain↑, ↑This act is pious↑ causes ↑This act is lovable by the gods↑ (*cf* Plato, *Euthyphro*, 10d), ↑This shape is a triangle↑ causes ↑The sum of the angles of this shape is straight↑ (*cf* Aristotle, *Posterior Analytics*, A, 2, 71b; 4, 73b–74a), ↑The One is perfect↑ causes ↑The Intellect exists↑ (*cf* Plotinus, *Enneades*, V, 1, 6–7).

Say that s is *unstable with respect to* t iff either [s causes t but, if the obtaining of s token-causes the obtaining of t, the s-[token-cause] does not coobtain with the t-[token-effect]][13] or [t causes s but, if the obtaining of t token-causes the obtaining of s, the s-[token-effect] contraobtains with the t-[token-cause]].[14] Say that s is *unstable** *with respect to* t iff, in addition to being unstable with respect to t, either [s causes t but s does not cause st] or [t causes s but $s|t$ does not cause s]. Say that s is *unstable*** *with respect to* t iff, in addition to being unstable* with respect to t, either [s causes t but, if the obtaining of s token-causes the obtaining of t, s does not coobtain with t][15] or [t causes s but, if the obtaining of t token-causes the obtaining of s, s contraobtains with t].[16] Say that a SA is *unstable* (*unstable**, *unstable***) iff it

[11] Throughout, for brevity, I typically use the former, less precise but more concise kind of construction in the text ('Some SAs cause some SAs'), but what I always mean is what is expressed by the latter, more precise but less concise kind ('The obtaining of some SAs causes the obtaining of some SAs', which is in turn obviously understood in the sense that the obtaining of s_0 and the obtaining of s_1 and the obtaining of s_2 ... and the obtaining of s_i—*i.e.* the obtaining of $s_0s_1s_2...s_i$—cause the obtaining of s_{i+1} and the obtaining of s_{i+2} and the obtaining of s_{i+3} ... and the obtaining of s_{i+j}—*i.e.* the obtaining of $s_{i+1}s_{i+2}s_{i+3}...s_{i+j}$—and in the sense that the obtaining of, say, ↑Snow is white↑ simply consists in snow's being white). Also, throughout, I use unqualified 'cause' and its relatives for *type* causation, while I always make it explicit when I mean *token* causation.

[12] Although the same tradition typically recognises also other kinds of natural causation (e.g. Aristotle, *Physics*, B, 3), for simplicity I'll ignore them and, by 'natural causation', really mean natural *efficient* causation. For what it's worth, also all the cases of nonnatural causation mentioned in this paper are cases of *efficient* causation (in the sense that they are cases where the cause is the *origin* of the effect).

[13] Throughout, I use square brackets to disambiguate constituent structure.

[14] The conditionals 'If the obtaining of s token-causes the obtaining of t ...' and 'If the obtaining of t token-causes the obtaining of s ...' are *not* supposed to be *vacuously true*: s's instability *leaves it open* that s obtains (and so, since a cause can only token-cause an effect if the former obtains, leaves it to that extent open that s token-causes t) and that s does not obtain (and so, since a cause can only token-cause an effect if the latter does not obtain, leaves it to that extent open that t token-causes s).

[15] Henceforth and until further notice (fn 51), for conciseness, I'll typically understand as implicit the proviso 'if the obtaining of s token-causes the obtaining of t'.

[16] While both the *combinational* and the *alternational* aspect of instability (instability*, instability**) and of related notions, as represented in each of the previous definitions in this paragraph by its two disjuncts respectively, will be mentioned in the most official formulations, to keep things simple I'll focus in the rest of this paper on the combinational aspect. Notice that, for kinds of causation over which complementation is *monotonic* (as is the case for the kind of causation focussed on in the paper), instability (instability*,

is unstable (unstable*, unstable**) with respect to some SA. Say that a SA is *stable* iff it is not unstable (and assume, plausibly, that, if s is stable, then, [if s causes t, s causes st] and, [if t causes s, $s|t$ causes s]). In typical cases, instability implies instability**, so that instability, instability* and instability** can be treated equivalently, and will in fact be so treated for most of this paper by simply talking about "instability". Yet, we'll see in Section 4 some cases where these three properties come apart, in which cases, for the purposes of the paper, the really crucial notion is that of instability (having noted that, in the further development of Section 5 the differences among instability, instability* and instability** will be largely overcome, see fn 59).

It might be natural to assume that *nonnatural* causation can only involve *stable* SAs: for example, ↑This act is pious↑ causes ↑This act is lovable by the gods↑, and coobtains with it. That is in sharp contrast with *natural* causation, which typically involves *unstable* causes: for example, w causes a, but w does not coobtain with a (it is impossible that both the wood is burning and only ashes remain).

As an aside, notice that the last kind of example, supposed to establish the instability of typical natural causes like w, is most naturally understood as involving SAs that are both *tensed* and *time-unspecific*[17] (as congenial to the *pre-Fregean* tradition, e.g. Aristotle, *Categories*, 5, 4a–b). For it is the coobtaining of such SAs that would seem most clearly to consist in their obtaining both *at the same time*, and, in turn, it is the obtaining at the same time by such SAs that is most clearly *apt to fail* in the kind of example in question. If that kind of example were however understood as involving *tenseless* or *time-specific* SAs, it would be much less clear that it would establish instability. For the example would now be dealing with SAs along the lines of, say, ↑At t_0, the wood is burning↑ and ↑At t_1, only ashes remain↑ (where t_0 and t_1 are appropriately related), and it is much less clear that the former SA does not coobtain with the latter SA. It is tensed and time-unspecific SAs that, as natural causes, are most clearly SAs that are typically unstable.[18]

Back from the aside, against the natural assumption mentioned in the second last paragraph it is arguable that nonnatural causation can also involve *unstable* SAs. For example, consider the conception of the natural-number system $\mathbb{N}$ as the *result of a development that, given 0 as a starting point, yields the other natural numbers by repeated applications of the operation of succession* (e.g. [8], §1).[19] Such conception

instability**) of s with respect to t in the combinational aspect is plausibly correlated with instability (instability*, instability**) of t^* with respect to s^* in the alternational aspect.

[17] I require both properties for the kind of SA I have in mind: ↑At (time) t, the wood was burning↑ is tensed but time-specific; conversely, ↑Two *plus* two is four↑ is time-unspecific but (very plausibly) tenseless.

[18] Thanks to Hannes Leitgeb for probing questions about these issues.

[19] The conception is, I take it, in itself plausible, and is one of the most natural ways of interpreting the basic elucidation of natural numbers as all and only those numbers comprising "*0 and whatever you can get to in finitely many steps*". Notice that, in the relevant sense, contrary to Dedekind's own spin, *there need not be anything mind dependent* about such development, just as there is nothing mind dependent about, say, the development of hypostases in neoplatonic philosophy. The example of the development of $\mathbb{N}$ is particularly suggestive in the context of this paper, since it arguably underlies the *extraction of* $\mathbb{N}$ *out of an arbitrary infinite set* in [9], §6, Definition 71, Theorem 72, Definition 73, where in turn the *existence of infinite sets* in the first place is established by an argument (§5, Theorem 66) closely related to the one given in [5], §13, which in turn relies on the infinite series of iterated truth attributions 'P', ''P' is true', '''P' is true' is true'… Bolzano explicitly remarks on the correspondence between that series and $\mathbb{N}$. For

can naturally be expressed in terms of SAs and causation. Let '$n_0n_1n_2\ldots n_i!$' be short for '↑Exactly natural numbers n_0, n_1, $n_2\ldots$ and n_i exist↑', and consider 0!. By succession, 0! causes 01!, but, since 0! is incompatible with 01! (as a consequence of the more general fact that certain objects' being the only Fs is incompatible with any other object's being F), 0!^01! is impossible, and so 0! does not coobtain with 01!.[20] Therefore, on this conception, 0! is unstable. Indeed, by analogous arguments, on this conception, for every i, $012\ldots i!$ is unstable.

One might observe that the impossibility of 0!^01! is due to the fact that its first argument is ↑Exactly 0 exists↑, and object that the development of $\mathbb{N}$ might at least equally well take the initial cause to be ↑0 exists↑ (and the initial effect to be ↑0 and 1 exist↑). But such SAs as ↑0, 1, 2… and i exist↑ would not be adequate inputs or outputs of succession. For succession is an operation that extends a *specific* initial segment of $\mathbb{N}$ by a *specific* extent: what succession does is to take—not any old natural number, but—what is the *limit* of the development of $\mathbb{N}$ for then *going beyond it*—not to any old distance, but—*one single step*. Making essentially the same point in a clearer if less direct way, ↑0 exists↑ cannot be the input of succession *in taking 0 and producing 1*, for that SA also obtains at a stage of the development of $\mathbb{N}$ where, say, 37 exists, and so at a stage where succession does *not* take 0 to produce 1, with the consequence that, if ↑0 exists↑ were such input, at some stage succession would apply to it while at some other stage, although that SA does obtain at that stage and succession does apply to that stage in general, succession would weirdly fail to apply to that SA in particular;[21] similarly, ↑0 and 1 exist↑ cannot be the output of succession *in taking 0 and producing 1*, for that SA also obtains at a stage where

both thinkers, broadly semantic notions and naive ascent principles governing them were the gateway to the existence of the actual infinite.

[20] It might be claimed that, since all natural numbers necessarily exist, both 0! and 01! are impossible too (and, indeed, for every i, $012\ldots i!$ is impossible too). However, in the case of systems resulting from such and similar developments, it is important to distinguish between what is *possible given only a partial development of the system* and what is *possible given the total development of the system* (just as it is important to distinguish between what is *possible at a certain time* and what is *possible at a later time*: for example, before eating breakfast it is possible that I'll fast for the day, while after eating breakfast it is not possible that I'll fast for the day), and it is solely lack of possibility of a SA given only the development of the system up to a certain point that prevents the SA from obtaining if the causation leading to the next point occurs. A broadly similar distinction must be drawn between what *obtains given only a partial development of the system* and what *obtains given the total development of the system*. Therefore, while, *given the total development of* $\mathbb{N}$, 0! and 01! are indeed not possible, *given only the development of* $\mathbb{N}$ *up to 0* 0! is possible (since, given only the development of $\mathbb{N}$ up to 0, it obtains) and so is 01! (since, given only the development of $\mathbb{N}$ up to 0, it is caused by a SA—*i.e.* 0!—which obtains), whereas 0!^01! is no more possible given only the development of $\mathbb{N}$ up to 0 than it is given the total development of $\mathbb{N}$. Throughout, I leave it to context to disambiguate whether, by certain unqualified occurrences of 'possible' and its relatives, I mean possibility given only the development of the system up to the relevant point or possibility given the total development of the system. Thanks to an anonymous referee for raising this issue.

[21] Don't say that it wouldn't be weird because succession only applies to ↑0 exists↑ under the condition that 0 is the greatest natural number that exists at the stage in question. For that would seem just a fancy way of saying that succession really applies to ↑0 exists and is the greatest natural number that exists↑, which is virtually the same SA as ↑Exactly 0 exists↑.

37 exists, and so at a stage where succession does *not* produce 1 by taking 0, with the consequence that, if ↑0 and 1 exist↑ were such output, at some stage succession would deliver it while at some other stage, although that SA does obtain at that stage and succession does deliver that stage in general, succession would weirdly fail to deliver that SA in particular.[22,23,24]

While plausible, such conception of $\mathbb{N}$ is of course not uncontentious: leaving it open whether it is *correct*, its being (beyond any serious doubt) *intelligible* suffices to show that, contrary to what might be natural to assume (given also the motivating example of instability given in the fourth last paragraph), *instability need not involve temporality* (which is also evidenced by the fact that *absolutely no temporal concept was used* in the characterisation of instability given in the fifth last paragraph). It would be an unbelievably gross and inadequate understanding of the development of $\mathbb{N}$ to think that, *at some time*, 0! obtains and, *at some other time*, 01! obtains. Obviously, the two SAs are not supposed to obtain at different *times*, but, to appeal now more explicitly to a crucial concept that already naturally emerged in the last paragraph, at different *stages* of the *atemporal* development of $\mathbb{N}$. A collection of atemporal objects not unusually comes with some *structure*, and in some cases such structure is both such as to *induce a wellorder* and such as to *reflect some underlying nonnatural causation*. Ordinary and scientific thought does employ a concept of stage that applies in precisely such cases, and that concept has proven extremely useful in understanding the *organisation principles* of such structures (think for example of the use of this concept made e.g. by [36] in understanding the organisation principles of the intended model of ZF-like set theories). While natural unstable causes typically obtain at different times than their relevant effects, nonnatural unstable causes typically obtain at different stages than their relevant effects.

It should by now be clear why, contrary to a nowadays influent, *Bolzanian* (fn 29) trend in metaphysics (e.g. [7]), I'm not using '*ground*' and its relatives to express nonnatural causation. For it is extremely plausible that *grounds are stable causes*: what grounds is the *foundation* on which what is grounded *rests*, and such founding extremely plausibly requires the *coexistence* of its *relata*. But, as I've argued in the last three paragraphs, not all nonnatural causes need be stable.[25] In causing its effects,

[22] Don't say that it wouldn't be weird because succession only delivers ↑0 and 1 exist↑ under the condition that 1 is the greatest natural number that exists at the stage in question. For that would seem just a fancy way of saying that succession really delivers ↑0 and 1 exist and 1 is the greatest natural number that exists↑, which is virtually the same SA as ↑Exactly 0 and 1 exist↑.

[23] Notice that, if you were particularly convinced either by the argument about input or by the one about output, you should be convinced of the conclusion of the other argument in virtue of the very plausible principle that the input and output of succession should be SAs of the same kind.

[24] Thanks to Hannes Leitgeb, Dan López de Sa, Sven Rosenkranz and Jeremy Wyatt for their constructive incredulity regarding these ideas.

[25] In this respect, from the perspective of this paper, the different use of 'ground' and its relatives made in the literature on the semantic paradoxes (for which [19] is an early reference) would seem to some extent more appropriate. I hasten to add, though, that, even on that use, there would still be some unstable SAs that would get counted as "grounded" (fns 59, 64) and, sort of conversely, some not unstable* SAs that would not get counted as "grounded" (fn 59).

any unstable cause, far from *founding*, *founders* instead.[26] Unstable causation is not a matter of *foundation*, but of *evolution*.

4 Causation

With this much by way of background, let's now proceed to develop a *theory* (*i.e.* set of sentences) $\mathfrak{C}$ of *nonnatural truth-related causation* (the one exemplified by ↑Snow is white↑ causing ↑'Snow is white' is true↑).[27] $\mathfrak{C}$ contains two kinds of principles: *causation principles*—saying that *certain causes cause certain effects*—and *metacausation principles*—saying that, *if certain causes cause certain effects, certain possibly other causes cause certain possibly other effects*. In other words, while causation principles are *inter[cause-effect]* principles, metacausation principles are *intracause* or *intraeffect* principles. $\mathfrak{C}$'s causation principles will concern *truth*; $\mathfrak{C}$'s metacausation principles will concern *the operations on SAs*. In the development of $\mathfrak{C}$ (as well as in those of the other theories in Sections 5 and 6), I'll assume, as usual, that the background logic (and mathematics) is classical,[28] since, even when theorising about a nonclassical logic with a philosophical application to a certain area justified by the area's sensitivity to certain nonclassical factors, it is typically compelling to assume that *whether something is valid in the logic is not equally sensitive to those factors*, and so that it remains a classical matter. The assumption becomes less compelling *once the logic is so strengthened that whether something is valid in the logic becomes sensitive to the same nonclassical factors that justify the logic's application to the relevant area* (as is done in the case of the semantic paradoxes e.g. by [46]); such strengthening does relate to some of the themes of this paper, but its proper treatment lies beyond its scope (see however fn 65 for some illustrative discussion).[29]

Before proceeding, some notes on notation. In the semiformal parts of this paper, I'll use $\frown$ to express nonnatural truth-related causation. However, it'll prove important to be able also to talk about the same principles understood as applying to other

[26]The pun is inspired by [18], I, 2, C, 3 (who masterfully plays in several ways with the German pair *Grund*/*zu Grunde*), an author to whom, as *per* the subtitle of this paper, the views I'm presenting are also indebted in more substantial respects that go however beyond the scope of the paper.

[27]Henceforth, 'cause' and its relatives should typically be understood as so qualified.

[28]The assumption is almost always made, but typically either on the not very compelling ground that it is *technically useful* or on the not very compelling ground that it is *pedagogically useful*. Pragmatism about *logic* is no better than pragmatism about *truth*.

[29]Formal theories of *causation in general* (e.g. [26]) typically focus on *natural* causation, so that their relevance for $\mathfrak{C}$ is somewhat limited; formal theories of *nonnatural causation in general* (e.g. [4], II, 3–5, III, §§162–222) typically focus on *stable* causation, so that, again, their relevance for $\mathfrak{C}$ is somewhat limited. Still, at least one point of comparison between both these kinds of theories on the one hand and $\mathfrak{C}$ on the other hand emerges clearly: both these kinds of theories typically assume that the relation of causation is *asymmetric* and *irreflexive*, whereas, as I'll make explicit in this section and Section 5 respectively, $\mathfrak{C}$ implies that the relation of nonnatural truth-related causation is in some cases symmetric and in some cases reflexive (as typically acknowledged to some extent or other by theories of *nonnatural truth-related causation*, e.g. [39], p. 130). Thanks to an anonymous referee for recommending a comparison of $\mathfrak{C}$ with extant formal theories of causation.

kinds of causation; moreover, in Section 5, we'll consider another prominent "arrow-ish" relation SAs can enter into, implication, and it'll prove equally important to be able also to talk about the same principles understood as applying to implication rather than to causation. For these reasons, those principles are typically more neutrally stated in terms of a generic $\rightsquigarrow$ (and their various versions concerning specific "arrow-ish" relations disambiguated—when they need be—by superscripting them with the intended arrow). Also, given a pair of subscripted principles, I'll use the unsubscripted name to refer to the conjunction of the two principles.

Quite properly, let's start with **p**. As *per* Section 2, **p** is the *absolutely positive* SA, *i.e.* the SA that *combines with every SA that obtains*.[30] Thus, $\mathfrak{C}$ contains the metacausation principles of *positivity*:

($\text{POS}_{\leq}$) If $s \rightsquigarrow t\mathbf{p}$ holds, $s \rightsquigarrow t$ holds, and, if $s \rightsquigarrow t$ holds, $s\mathbf{p} \rightsquigarrow t$ holds;[31]
($\text{POS}_{\geq}$) If $s \rightsquigarrow t$ holds, $s \rightsquigarrow t\mathbf{p}$ holds, and, if $s\mathbf{p} \rightsquigarrow t$ holds, $s \rightsquigarrow t$ holds.

In other words, **p** is a *positive nothing* (*nihil privativum*), what is *always ruled in*.

But how can **p** be understood in terms of the other operations? The answer lies in complementation and alternation. Although no complementation-free SA may be guaranteed to obtain, if any s fails to obtain, complementation is *weak enough* to record such failure in the form of the obtaining of s^*, and alternation is *weak enough*

[30] *Even in causal contexts* (such strong understanding of claims of this kind is operative in this and the next three paragraphs). Thanks to David Ripley for discussion on the proper characterisation of **p** and **n**.

[31] Most of $\mathfrak{C}$'s metacausation principles (the exceptions being ($\text{STA}^{\curvearrowright}$) and ($\text{JUXT}^{\curvearrowright}$) below in the text) can be seen as pairs one of whose elements (labelled with subscript $\leq$) is essentially to the effect that a SA s resulting from a certain series of operations has a *causal role at least as strong as* a SA t and whose other element (labelled with subscript $\geq$) is essentially to the effect that t has a *causal role at least as strong as* s; any such pair of principles can thus be seen as *fixing the causal strength* of a SA resulting from the series of operations in question. In this respect, most of $\mathfrak{C}$'s metacausation principles are *similar* to the *operational metarules of a standard sequent calculus* [13]. In addition to the existence of the exceptions noted above (not to speak of the *causation*—rather than *metacausation*—principle ($\text{A}^{\curvearrowright}$) below in the text), there are however several further respects of *dissimilarity*. Firstly, each of the metacausation principles in question determines that s (t) has a causal role at least as strong as t (s) by determining *both* that, whenever s (t) is an effect, so is t (s), *and* that, whenever t (s) is a cause, so is s (t). Importantly, the former, *suffixing* clause is *independent* from the latter, *prefixing* clause; they would be *equivalent* if $\curvearrowright$ were *reflexive* and *transitive* (as the derivability relation of a standard sequent calculus is), but it is very dubious that $\curvearrowright$ has either of these properties (and, in fact, we'll see in this section and Section 5 that it is important for the purposes of this paper that it has neither). Secondly, most of the metacausation principles in question (the exceptions being ($\text{POS}^{\curvearrowright}$) and ($\text{NEG}^{\curvearrowright}$) below in the text) fix the causal strength of a SA resulting from a certain series of *at least two* operations relative to a SA *resulting from another series of operations*, whereas in a standard sequent calculus the logical strength of a sentence with a certain principal operator *but arbitrary nonprincipal operators* is fixed in *purely structural* terms. (More in detail, those metacausation principles essentially determine the *behaviour of each composition operation when taking complementation or the other composition operation as argument*, and basically amount to saying that *causal reality is exhaustive, exclusive and distributive*: while those properties are plausibly fundamental at the level of causal reality, for better or worse they are not taken as fundamental in proof theory as typically practised.) Thirdly, $\mathfrak{C}$ totally abolishes the invidious distinction drawn by standard sequent calculi between *operators* on the one hand and (premise- and conclusion-) *aggregators* on the other hand: we can only aggregate different SAs as causes or effects by building up more complex SAs with the two composition operations. (Similar comments apply concerning similarities and dissimilarities between the theory of truth-related implication extracted from $\mathfrak{C}$ in Section 5 and the operational metarules of a standard sequent calculus.) Thanks to an anonymous referee for comments that prompted an elaboration of the material in this fn.

to record such *positive* dependence, so that $s|s^*$ is guaranteed to obtain. And any such basic fact about a SA combines with every SA that obtains (every SA that obtains combines with such possibility concerning each SA). Thus, $\mathfrak{C}$ contains the metacausation principles of *exhaustion*:

($\text{EXH}_{\leq}$) If $s \rightsquigarrow t(u|u^*)$ holds, $s \rightsquigarrow t\mathbf{p}$ holds, and, if $s\mathbf{p} \rightsquigarrow t$ holds, $s(u|u^*) \rightsquigarrow t$ holds;
($\text{EXH}_{\geq}$) If $s \rightsquigarrow t\mathbf{p}$ holds, $s \rightsquigarrow t(u|u^*)$ holds, and, if $s(u|u^*) \rightsquigarrow t$ holds, $s\mathbf{p} \rightsquigarrow t$ holds.

Conversely, $\mathbf{n}$ is the *absolutely negative* SA, *i.e.* the SA that *alternates with only SAs that obtain*. Thus, $\mathfrak{C}$ contains the metacausation principles of *negativity*:

($\text{NEG}_{\leq}$) If $s \rightsquigarrow t|\mathbf{n}$ holds, $s \rightsquigarrow t$ holds, and, if $s \rightsquigarrow t$ holds, $s|\mathbf{n} \rightsquigarrow t$ holds;
($\text{NEG}_{\geq}$) If $s \rightsquigarrow t$ holds, $s \rightsquigarrow t|\mathbf{n}$ holds, and, if $s|\mathbf{n} \rightsquigarrow t$ holds, $s \rightsquigarrow t$ holds.

In other words, $\mathbf{n}$ is a negative nothing (*nihil negativum*), what is *always ruled out*.

But how can $\mathbf{n}$ be understood in terms of the other operations? The answer lies in complementation and combination. Although no complementation-free SA may be guaranteed to fail to obtain, if any s obtains, complementation is *strong enough* to record such obtaining in the form of failure of s^* to obtain, and combination is *strong enough* to record such *negative* dependence, so that ss^* is guaranteed to fail to obtain. And any such basic nonfact about a SA alternates only with SAs that obtain (only SAs that obtain alternate with such impossibility concerning each SA). Thus, $\mathfrak{C}$ contains the metacausation principles of *exclusion*:

($\text{EXC}_{\leq}$) If $s \rightsquigarrow t|uu^*$ holds, $s \rightsquigarrow t|\mathbf{n}$ holds, and, if $s|\mathbf{n} \rightsquigarrow t$ holds, $s|uu^* \rightsquigarrow t$ holds;
($\text{EXC}_{\geq}$) If $s \rightsquigarrow t|\mathbf{n}$ holds, $s \rightsquigarrow t|uu^*$ holds, and, if $s|uu^* \rightsquigarrow t$ holds, $s|\mathbf{n} \rightsquigarrow t$ holds.

Time for truth. Let's assume that, for every s, there is a designated sentence $\mathsf{exp}(s)$ of the language that expresses s, and let $s^T = \uparrow\mathsf{exp}(s)\text{ is true}\uparrow$. Now, *truth is correspondence with the facts*,[32] and, keeping fixed what sentences represent, *whether they correspond with the facts is determined by which facts there are*: therefore, *the obtaining of a corresponded-with SA* causes the *truth* of a sentence, while *failure of a corresponded-with SA to obtain* causes the *untruth* of a sentence. Thus, $\mathfrak{C}$ contains the causation principles of *positive ascent*:

(A_P) $s \rightsquigarrow s^T$ holds

[32] Very roughly, in the sense that "the speech that speaks of beings as they are is true" (Plato, *Cratylus*, 385b). I believe that the doctrine of truth as correspondence is essentially correct, but that it also falls dramatically short of vindicating any general principle—so prominent in the alternative tradition of *deflationism* (e.g. [10, 22])—correlating its being the case that P with 'P' 's being true [41, 43, 49]. Nevertheless, as we'll see, truth as correspondence suffices to vindicate principles strong enough to generate semantic paradoxes no less than deflationary truth does. Thanks to an anonymous referee for feedback on these issues.

and of *negative ascent*:

(A_N) $s^* \rightsquigarrow s^{T^*}$ holds.

Some theories of truth contain instead as basic the principle of *positive descent*:

(D_P) $s^T \rightsquigarrow s$

(typically understood not as a causation principle). Notice that (D_P) follows from (A_N) by the metacausation principle of *contraposition*:

(CONTRAP) If $s \rightsquigarrow t$ holds, $t^* \rightsquigarrow s^*$ holds.

However, it is in the essence of causation not to satisfy (CONTRAP): quite generally, letting $\rightsquigarrow$ express a kind of causation, that $s \rightsquigarrow t$ holds does not imply that $t^* \rightsquigarrow s^*$ holds.[33] For example, letting $\rightsquigarrow$ express natural causation, that $w \rightsquigarrow a$ holds does not imply that $a^* \rightsquigarrow w^*$ holds: what causes the wood not to be burning are such things as humidity, wind, absence of sparks *etc.* rather than the mere existence of some part or other of the wood which is not ashes. More in particular, in our context, that $s \curvearrowright t$ holds does not imply that $t^* \curvearrowright s^*$ holds, and so (CONTRAP$^{\curvearrowright}$) does not hold.[34]

Moreover, not only does ($D_P^{\curvearrowright}$) *not follow* from ($A_N^{\curvearrowright}$); it is anyways *in itself implausible*. For a fundamental feature of truth is that *truth is nonsymmetrically caused by reality*, in the sense that:

($R^{\curvearrowright}T_\forall$) For every s, if s^T obtains, the obtaining of s token-causes the obtaining of s^T, and, if s^{T^*} obtains, the obtaining of s^* token-causes the obtaining of s^{T^*}

holds whereas:

($T^{\curvearrowright}R_\forall$) For every s, if s obtains, the obtaining of s^T token-causes the obtaining of s, and, if s^* obtains, the obtaining of s^{T^*} token-causes the obtaining of s^*

does not hold (e.g. Aristotle, *Categories*, 12, 14b, who would however seem to be overstating his case into a negation of ($T^{\curvearrowright}R_\exists$) below). Now, if the first component of ($T^{\curvearrowright}R_\forall$) does not hold, it presumably follows that, for some s, s^T obtains but the obtaining of s^T does not token-cause the obtaining of s, which in turn presumably entails that ($D_P^{\curvearrowright}$) does not hold. And, even if either of those two presumed entailments should somehow fail, it remains the case that the most natural reasons for thinking that ($T^{\curvearrowright}R_\forall$) does not hold are just as good reasons for thinking that ($D_P^{\curvearrowright}$) does not hold. For example, ($T^{\curvearrowright}R_\forall$) does not hold because, evidently, $\uparrow$Snow is

[33] At the very very best, for every kind of causation over which complementation is monotonic (as is the case, by the connection between (A_P) and (A_N), for the kind of causation focussed on in this paper), that $s \rightsquigarrow t$ holds implies that $s^* \rightsquigarrow t^*$ holds.

[34] Since (CONTRAP) arguably holds for implication, we have here a first example of the *divergence* between causation and implication. Still, we'll see in Section 5 that from $\mathfrak{C}$ we can *extract* a theory of truth-related implication, in the sense that the mere addition of very few very general and very compelling principles that are distinctive of implication *vis-à-vis* causation suffices to *recover* the whole wealth of specific desirable principles for implication (as for the recovery of (CONTRAP) in particular, see fn 50). (Conversely, we'll further see in Section 5 that the same addition also suffices to *wreck* one principle ((DISTR) below in the text) which arguably holds for causation.) Thanks to an anonymous referee for feedback on (CONTRAP).

white↑ obtains but the obtaining of ↑Snow is white↑^{T} does not token-cause the obtaining of ↑Snow is white↑; but that is just as good a counterexample to ($D_P^{\curvearrowright}$) too. Analogous comments apply to ($D_N^{\curvearrowright}$). (On the other hand, I note that, although logically independent, the conjunction of ($R^{\curvearrowright}T_\forall$) and the negation of ($T^{\curvearrowright}R_\forall$) is in great harmony with ($A^{\curvearrowright}$).)

Let's add something a bit more exciting. One might have thought that the non-symmetric causation of truth by reality includes a claim stronger than the negation of ($T^{\curvearrowright}R_\forall$), namely the negation of:

($T^{\curvearrowright}R_\exists$) For some s, if s obtains, the obtaining of s^T token-causes the obtaining of s, and, if s^* obtains, the obtaining of s^{T^*} token-causes the obtaining of s^*.

But ($T^{\curvearrowright}R_\exists$) *should not be denied, since it can virtually be proven given selfreferential ascending truth* (*i.e.* truth with selfreference and (A)). Suppose that there is a SA l of its being the case that the designated sentence of the language that expresses l is not true, so that $l = \uparrow\mathsf{exp}(l)$ is not true↑. Then, by the plausible *connection between negation and complementation* already remarked on in Section 2:

(NEGCOMP) ↑It is not the case that $P\uparrow = \uparrow P\uparrow^*$,

$l = \uparrow\mathsf{exp}(l)$ is not true↑ $= \uparrow\mathsf{exp}(l)$ is true↑$^* = l^{T^*}$ (and so, by involutivity, $l^* = l^T$). Now, it is extremely plausible that, if l^{T^*} obtains, the obtaining of l^* token-causes the obtaining of l^{T^*} (that is after all entailed by ($R^{\curvearrowright}T_\forall$) and in great harmony with ($A_N^{\curvearrowright}$)). Taking l as witness, since $l^{T^*} = l$ and $l^* = l^T$, that gives us the first conjunct of ($T^{\curvearrowright}R_\exists$). Moreover, it is extremely plausible that, if l^T obtains, the obtaining of l token-causes the obtaining of l^T (that is after all entailed by ($R^{\curvearrowright}T_\forall$) and in great harmony with ($A_P^{\curvearrowright}$)). Taking again l as witness, since $l^T = l^*$ and $l = l^{T^*}$, that gives us the second conjunct of ($T^{\curvearrowright}R_\exists$), and so ($T^{\curvearrowright}R_\exists$) follows. Although a robust version of *alethic realism* is correct to the effect that, typically, reality causes truth but not *vice versa*, a mild version of *alethic idealism* is also correct to the effect that, sometimes, truth does cause reality.[35]

Now that selfreferential ascending truth has been introduced, I can finally declare which SAs are unstable for the purposes of this paper: *all and only those expressed by sentences that involve a selfreferential attribution of truth* (under an appropriately

[35] Without naming names, it is shocking to see the contemporary literature on *grounding and truth making* treating as a mantra a principle like:

($R^{\mapsto}T_\forall$) For every s, if s^T obtains, the obtaining of s token-grounds the obtaining of s^T, and, if s^{T^*} obtains, the obtaining of s^* token-grounds the obtaining of s^{T^*}

(where $\mapsto$ expresses grounding), which, in classical logic (which typically goes unchallenged in that literature), can virtually be disproven given selfreferential truth. Suppose for *reductio ad absurdum* that l^T obtains. Then, by ($R^{\mapsto}T_\forall$), the obtaining of l token-grounds the obtaining of l^T, and so, by stability of grounds, ll^T—that is, $l^{T^*}l^T$—obtains, which is impossible. Therefore, by *reductio ad absurdum*, l^T does not obtain, and so l^{T^*} obtains. Then, by ($R^{\mapsto}T_\forall$), the obtaining of l^* token-grounds the obtaining of l^{T^*}, and so, by stability of grounds, $l^*l^{T^*}$—that is, $l^Tl^{T^*}$—obtains, which is impossible. (Although shocking, the attitude is by no means exceptional: without naming names, compare the analogous attitude towards the (T)-schema in the contemporary literature on *alethic pluralism*.) Notice that, even in classical logic, ($R^{\curvearrowright}T_\forall$) does not suffer from the same problem if, as I'm assuming in this paper, contrary to grounding, nonnatural truth-related causation can be unstable.

broad understanding of involvement, which includes involving an attribution of truth *to such sentences*, and under an appropriately broad understanding of selfreference, which includes *nonwellfounded referential chains* of all sorts).[36] (Such characterisation of instability might seem to *overgenerate*, but it arguably does not, since the usual understanding of semantic paradoxicality arguably *undergenerates*, see fn 64.) I can also provide the fundamental witness to their instability: if s is some such SA, by $(\mathrm{A}^{\curvearrowright}_{\mathrm{P}})$, it causes s^T, and *that is the fundamental effect with which it does not coobtain* (all other such effects like e.g. $s^T\mathbf{p}$ being derivative on it). Instability arises from selfreferential ascent. Thus, to take the paradigmatic example of l, by $(\mathrm{A}^{\curvearrowright}_{\mathrm{P}})$, l causes l^T, and that is the fundamental effect with which it does not coobtain.[37] Since $\mathfrak{C}$ is so being constructed as to allow for the instability of most SAs, the consequences of the stability of a SA must explicitly be built into $\mathfrak{C}$. Thus, $\mathfrak{C}$ contains the metacausation principle of *stability*:

(STA) For every STABLE s, if $s \rightsquigarrow t$ holds, $s \rightsquigarrow st$ holds, and, if $t \rightsquigarrow s$ holds, $s|t \rightsquigarrow s$ holds.[38]

It's time to move on to the rest of the metacausation principles. One reason for denying the metacausation principle of *monotonicity*:

(MON) If $s \rightsquigarrow t$ holds, $su \rightsquigarrow t$ holds, and, if $s \rightsquigarrow t$ holds, $s \rightsquigarrow t|u$ holds

is that natural causes do not *necessitate* their effects, and are indeed *defeasible*. For example, w causes a, but does not necessitate it, and the causation is indeed defeated by e.g. ↑The wood is constantly being reconstituted↑. It is true that, since nonnatural truth-related causes do necessitate their effects, *that* reason for denying (MON) drops in our context. But (MON) does not hold even in the absence of causal defeat, *simply because it is often a combinandum rather than a combination that is the cause, and it is often an alternandum rather than an alternation that is the effect*. For example, ↑Snow is white↑↷↑Snow is white↑^{T} holds, but neither ↑Snow is white↑ ˆ ↑Grass is green↑↷↑Snow is white↑^{T} nor ↑Snow is white↑↷↑Snow is white↑^{T} | ↑Grass is green↑ do.

[36] In this paper, I leave this at the status of a reasonable working assumption. I hope I'll be able in future work to vindicate such assumption.

[37] Notice that *not everyone who accepts* $(\mathrm{A}^{\curvearrowright}_{\mathrm{P}})$ *is committed to l's being unstable*. It is at least coherent to accept that $(\mathrm{A}^{\curvearrowright})$ holds and l (and, by the same token, l^*) is stable. On this general kind of view, there are then essentially two more specific options. On the first option, recasting *dialetheic* theories (e.g. [1, 30]) in terms of causation, since one accepts that each of l and l^* causes ll^* (and accepts that either l obtains or l^* obtains), one accepts that ll^* obtains. On the second option, recasting *supervaluationist* and *antialetheic* theories (e.g. [28] for the former and [6, 11] for the latter) in terms of causation, since one accepts that each of l and l^* causes ll^* (and rejects that ll^* obtains), one rejects both that l obtains and that l^* obtains.

[38] The intended interpretation of 'STABLE' is the property of being stable, but, since we'll be doing some model theory of (STA) at the end of this section and some regimentation of it in Section 6, it helps clarity to use a different, dedicated expression. Also, notice that, as *per* Section 3, $(\mathrm{STA}^{\curvearrowright})$ really captures *a consequence of* stability rather than stability *itself*. But that should not be surprising: while $\mathfrak{C}$, a theory of *type* causation, has an ideal format for the purposes of this paper, that is not a format where the *token*-centred extra strength of stability is easily manifested (moreover, in the further development of Section 5, such kind of token-centred difference will be largely overcome, see fn 59).

Still, even if (MON$^\frown$) fails in the attempt at expanding an *individual* causal claim into a *composite* one, there must be a way of extracting from individual causal claims composite causal claims. Firstly, there must be a way of extracting from individual causal claims *combined* causal claims, in particular a causal claim concerning the effects (causes) of the combination of the individual causes (effects). In our context, *where causal defeat is absent*, a plausible such extraction is that *each of the combined causes still causes the relevant effect, the result being the combination of those effects*. Secondly, there must be a way of extracting from individual causal claims *alternated* causal claims, in particular a causal claim concerning the effects (causes) of the alternation of the individual causes (effects). In our context, *where again causal defeat is absent*, a plausible such extraction is that *each of the alternated causes still causes the relevant effect, the result being the alternation of those effects*. In this complex sense, *causations are preserved under compositions*. Thus, $\mathfrak{C}$ contains the metacausation principle of *juxtaposition*:

(JUXT) If $s \rightsquigarrow t$ and $u \rightsquigarrow v$ hold, $su \rightsquigarrow tv$ and $s|u \rightsquigarrow t|v$ hold.

While (JUXT) yields composite causal claims, a principle is now required *to get the composed SAs to interact with one other*. Classical logic—along with many other nonclassical logics (intuitionist, many-valued, relevant *etc.*)—suggests the metacausation principles of *distribution*:

(DISTR$^\frown_{|\leq}$) If $s \rightsquigarrow t(u|v)$ holds, $s \rightsquigarrow tu|tv$ holds, and, if $st|su \rightsquigarrow v$ holds, $s(t|u) \rightsquigarrow v$ holds;

(DISTR$^\frown_{|\geq}$) If $s \rightsquigarrow tu|tv$ holds, $s \rightsquigarrow t(u|v)$ holds, and, if $s(t|u) \rightsquigarrow v$ holds, $st|su \rightsquigarrow v$ holds;

(DISTR$_{|^\frown\leq}$) If $s \rightsquigarrow t|uv$ holds, $s \rightsquigarrow (t|u)(t|v)$ holds, and, if $(s|t)(s|u) \rightsquigarrow v$ holds, $s|tu \rightsquigarrow v$ holds;

(DISTR$_{|^\frown\geq}$) If $s \rightsquigarrow (t|u)(t|v)$ holds, $s \rightsquigarrow t|uv$ holds, and, if $s|tu \rightsquigarrow v$ holds, $(s|t)(s|u) \rightsquigarrow v$ holds.

One might think that, because of instability, (DISTR) does not hold even for natural *common-or-garden* causation.[39] For example, letting $\rightsquigarrow$ express natural causation and $e = \uparrow$The wood exists$\uparrow$, (DISTR$^\frown_{|\leq}$) entails that, if $s \rightsquigarrow w(e|e^*)$ holds, $s \rightsquigarrow we|we^*$ holds. However, while $w(e|e^*)$ is unproblematic, we^* is impossible, and one might think that so is we. we would indeed be impossible given the metacausation principle of *separation*:

(SEP) If $s \rightsquigarrow t$ holds, $su \rightsquigarrow tu$ and $s|u \rightsquigarrow t|u$ hold,

since, given that $w \rightsquigarrow a$ holds, by (SEP) $we \rightsquigarrow ae$ would hold, and ae is impossible. But, contrary to (JUXT), (SEP) is not a plausible principle *of causation*: causation acts on a composition *as a whole* rather than, *separating* the two components, acting on one while leaving the other one alone. If only (JUXT) rather than (SEP) is available, we only get that, for some s such that $e \rightsquigarrow s$ holds, $we \rightsquigarrow as$ holds,

[39] I'll set aside in this paper the moot issue of whether it holds for natural *quantum-mechanical* causation (see (Zardini, E., Against the world, unpublished) for some discussion).

and there is no reason to think that *as* is impossible. Generalising from the features emerged in this discussion, it is plausible to expect that (DISTR) for *natural* causation is compatible with instability. Even more generally, it is plausible to expect that (DISTR) for causation *in general* is compatible with instability.

For the remainder of this section, I want to defend the latter plausible expectation by zooming in on what is for the purposes of this paper a particularly prominent consequence of (DISTR). To wit, an important feature of (DISTR) is that it implies the metacausation principles of *contraction*:

(CONTR$_{\cdot}$) If $s \rightsquigarrow t$ holds, $s \rightsquigarrow tt$ holds, and, if $ss \rightsquigarrow t$ holds, $s \rightsquigarrow t$ holds;
(CONTR$_{|}$) If $s \rightsquigarrow t|t$ holds, $s \rightsquigarrow t$ holds, and, if $s \rightsquigarrow t$ holds, $s|s \rightsquigarrow t$ holds.[40]

However, (CONTR$^{\curvearrowright}$) *is arguably unproblematic*. Let's consider for example a paradigmatically unstable SA like l, for which one might think that the alleged problematicity of (CONTR$^{\curvearrowright}$) emerges (focussing on (CONTR$^{\curvearrowright}_{\cdot}$)). For every s such that $s \curvearrowright l$ holds, by (CONTR$^{\curvearrowright}_{\cdot}$) $s \curvearrowright ll$ does hold, but that in turn only implies, by (A$^{\curvearrowright}_{\mathrm{P}}$), (JUXT$^{\curvearrowright}$) and (A$^{\curvearrowright}_{\mathrm{N}}$), that the causal chain $s \curvearrowright ll \curvearrowright l^T l^T \curvearrowright ll \ldots$ (as well as $s \curvearrowright lll \curvearrowright l^T l^T l^T \curvearrowright lll \ldots$, $s \curvearrowright llll \curvearrowright l^T l^T l^T l^T \curvearrowright llll \ldots$ *etc.*) holds.

I'd like to buttress with more general grounds the claim that (CONTR$^{\curvearrowright}$) is unproblematic, addressing what I take is the main worry about its compatibility with instability (focussing again on (CONTR$^{\curvearrowright}_{\cdot}$)). If a SA b_0 is unstable, a logician on loan to the theory of causation (for example, [42])—who is pulled towards reading arrows as some sort or other of implication—would expect that it causes not only a SA b_1 with respect to which it is unstable, but also a SA b_2 that, because of that instability, does not coobtain with b_1, with the effect that b_0 does not cause $b_1 b_2$. On the logician's expectation, we'd thus have a sort of *causal branching* where $b_0 \curvearrowright b_1$ and $b_0 \curvearrowright b_2$ hold, but $b_0 \curvearrowright b_1 b_2$ does not. Put more informally, although b_0 causes b_1 and causes b_2, and although it thus causes *either*, it does not cause *both*. However, it is in the essence of (JUXT$^{\curvearrowright}$) to entail that, whenever a cause causes either but not both effects, its selfcombination causes both. Therefore, although b_0 does not cause both b_1 and b_2, $b_0 b_0$ does. Put more formally, although $b_0 \curvearrowright b_1 b_2$ does not hold, $b_0 b_0 \curvearrowright b_1 b_2$ does. Against the superficial impression that the selfcombination of a SA cannot but boil down to the SA itself, the extremely plausible interpretation forced by (JUXT$^{\curvearrowright}$) of the causal strength of a combination makes the selfcombination of an unstable SA *in principle* causally much stronger than the SA itself. That is, I take it, the main worry about the compatibility between (CONTR$^{\curvearrowright}$) and instability.

However, I've emphasised 'in principle' in the second last sentence because, given both the nature of nonnatural truth-related causation and the tight constraints that

[40] Reason for (CONTR$_{\cdot}$): suppose that $s \rightsquigarrow t$ holds. Then, by (POS$_{\geq}$), $s \rightsquigarrow t\mathbf{p}$ holds, and so, by (EXH$_{\geq}$), $s \rightsquigarrow t(t|t^*)$ holds. By (DISTR$_{\cdot|\leq}$), $s \rightsquigarrow tt|tt^*$ holds, and so, by (EXC$_{\leq}$), $s \rightsquigarrow tt|\mathbf{n}$ holds, and hence, by (NEG$_{\leq}$), $s \rightsquigarrow tt$ holds. Suppose next that $ss \rightsquigarrow t$ holds. Then, by (NEG$_{\leq}$), $ss|\mathbf{n} \rightsquigarrow t$ holds, and so, by (EXC$_{\leq}$), $ss|ss^* \rightsquigarrow t$ holds. By (DISTR$_{\cdot|\leq}$), $s(s|s^*) \rightsquigarrow t$ holds, and so, by (EXH$_{\geq}$), $s\mathbf{p} \rightsquigarrow t$ holds, and hence, by (POS$_{\geq}$), $s \rightsquigarrow t$ holds. Reason for (CONTR$_{|}$): suppose that $s \rightsquigarrow t|t$ holds. Then, by (POS$_{\geq}$), $s \rightsquigarrow (t|t)\mathbf{p}$ holds, and so, by (EXH$_{\geq}$), $s \rightsquigarrow (t|t)(t|t^*)$ holds. By (DISTR$_{|\cdot\geq}$), $s \rightsquigarrow t|tt^*$ holds, and so, by (EXC$_{\leq}$), $s \rightsquigarrow t|\mathbf{n}$ holds, and hence, by (NEG$_{\leq}$), $s \rightsquigarrow t$ holds. Suppose next that $s \rightsquigarrow t$ holds. Then, by (NEG$_{\leq}$), $s|\mathbf{n} \rightsquigarrow t$ holds, and so, by (EXC$_{\leq}$), $s|ss^* \rightsquigarrow t$ holds. By (DISTR$_{|\cdot\geq}$), $(s|s)(s|s^*) \rightsquigarrow t$ holds, and so, by (EXH$_{\geq}$), $(s|s)\mathbf{p} \rightsquigarrow t$ holds, and hence, by (POS$_{\geq}$), $s|s \rightsquigarrow t$ holds.

$\mathfrak{C}$ puts on $\curvearrowright$, *as a matter of fact causal branching does not occur*, and so there is after all no good reason for thinking that the selfcombination of an unstable SA is *in fact* causally stronger than the SA itself. Firstly, and less conclusively, in the context of *undefeasible* causation (as nonnatural truth-related causation is), it is extremely unclear what it would be for a cause to cause *either* of two effects but *not both*. Indeed, it is extremely unclear what it would be for a possible cause to *cause either of two effects that do not coobtain*.[41] And, even granting that all that could be made reasonably clear (and plausible), it would then become utterly unclear why that does not constitute a sort of *quasi(self)defeasibility* calling for restrictions on (JUXT) just as well as normal defeasibility does.

Secondly, and more conclusively, it is natural and helpful to appeal to a crucial concept already emerged in Section 3 and understand nonnatural truth-related causation as proceeding by ordinal-indexed *stages of truth evaluation* (STEs),[42] so that, for every ordinal α, if s belongs to the STE $\mathsf{ste}(\alpha)$ and $s \curvearrowright t$ holds, t belongs to $\mathsf{ste}(\alpha + 1)$ (STEs stand to nonnatural truth-related causation as states of a physical system stand to natural causation). Extremely plausibly, s^* belongs to a STE iff s does not belong to it; st belongs to a STE if s belongs to it and t belongs to it; $s|t$ belongs to a STE only if s belongs to it or t belongs to it;[43] if s is stable and belongs to $\mathsf{ste}(\alpha)$, then, for every $\beta \geq \alpha$, s also belongs to $\mathsf{ste}(\beta)$. Then, recalling what I've said about instability in this section, instability arises only in cases of ascent, and so, recalling also what I've said about stages in Section 3, only between two SAs belonging, for some α, to $\mathsf{ste}(\alpha)$ and to $\mathsf{ste}(\alpha + 1)$ respectively. To take the paradigmatic example of l, we have that l, which, for some α, belongs to $\mathsf{ste}(\alpha)$, is unstable with respect to l^T, which belongs to $\mathsf{ste}(\alpha + 1)$ (since l does not coobtain with l^T, l itself does not belong to $\mathsf{ste}(\alpha + 1)$). Therefore, since in causal branching b_0 is supposed to be unstable with respect to b_1 and b_2 is supposed somehow to partake in that instability, we may assume that causal branching occurs only if, *although* $b_0 \curvearrowright b_2$ *holds*, for some α, both b_0 *and* b_2 belong to $\mathsf{ste}(\alpha)$ while b_1 belongs to $\mathsf{ste}(\alpha + 1)$.

But, if s is unstable, is $s \curvearrowright t$ holding compatible with t's belonging to the same STE as s? The logician does expect it to be such, since, because of her training, she backslides into understanding $\curvearrowright$ as some sort or other of *implication*, and implication can certainly hold between such SAs (if implication is in at least the relevant cases reflexive that will do, but we don't even need to go to those extremes, since s certainly implies $s|s^*$, which certainly belongs to every STE). But the logician's expectation is undermined precisely because $\curvearrowright$ expresses a certain kind of *causation* rather than

[41] I think that the situation dramatically changes in the case of defeasible causation. To sketch an example (which I discuss more extensively in [50], pp. 499–500; [52]), an invitation to a party sent to all of one's friends f_0, f_1, $f_2 \ldots$, $f_{1,000,000}$ causes *each*, and so *any*, of f_0, f_1, $f_2 \ldots$, $f_{1,000,000}$ to come, but does not cause *all* of f_0, f_1, $f_2 \ldots$, $f_{1,000,000}$ to come (where 'any' is the arbitrary-arity version of binary 'either' and 'all' the arbitrary-arity version of binary 'both')—for one thing, given such invitation, for every i, it is *extremely likely* that f_i will come, but it is *extremely unlikely* that, for every i, f_i will come.

[42] 'Truth evaluation' is really just another name for ascent, but 'stage of truth evaluation' has the advantage over 'stage of ascent' of generating, in this paper, an unambiguous acronym. STEs play an important role in [23] and in many subsequent approaches to the semantic paradoxes, especially the *revision-theoretic* one [2, 15, 16, 20, 21, 40].

[43] Notice that the previous clauses in the text only make sense in the presence of (CONTR$\curvearrowright$).

any sort of *implication*, and, in fact, $s \curvearrowright t$ holding is *typically* incompatible with t's belonging to the same STE as s. Unstable causes typically do not have intrastage effects. To take examples of the two kinds of cases just mentioned, neither $l \curvearrowright l$ nor $l \curvearrowright l|l^*$ hold (and $\mathfrak{C}$ has precisely been designed, among other things, to reflect that).

Now, all that does not mean that $s \curvearrowright t$ is *never* compatible with t's belonging to the same STE as s, and, in fact, the two things are *sometimes* compatible. Let $t_0 = \uparrow\mathsf{exp}(t_0)$ is true$\uparrow$ $\hat{}$ $\uparrow$Snow is white$\uparrow$, so that $t_0 = t_0^T$ $\hat{}$ $\uparrow$Snow is white$\uparrow$. By ($\mathrm{A}_{\mathrm{P}}^{\curvearrowright}$), $t_0 \curvearrowright t_0^T$ holds, but, since $t_0 = t_0^T$ $\hat{}$ $\uparrow$Snow is white$\uparrow$, for every α, if t_0^T and $\uparrow$Snow is white$\uparrow$ belong to $\mathsf{ste}(\alpha)$, so does t_0. Therefore, there are cases where $s \curvearrowright t$ holds while t belongs to the same STE as s. However, *by their very nature*, such cases are not cases where, assuming that s obtains, there is reason for thinking that t does not coobtain with some other effect of s. For example, two straightforward effects of t_0 are t_0^T (by ($\mathrm{A}_{\mathrm{P}}^{\curvearrowright}$)) and $t_0^{T^T}$ $\hat{}$ $\uparrow$Snow is white$\uparrow^T$ (by ($\mathrm{A}_{\mathrm{P}}^{\curvearrowright}$) and (JUXT$^{\curvearrowright}$)), and, assuming that t_0 obtains, there is obviously no reason for thinking that t_0^T does not coobtain with them.

This is an appropriate moment to make a digression into the differences among instability, instability* and instability**. Notice first that the case discussed in the last paragraph provides an example of a SA—t_0—that is unstable* and yet, since it coobtains with the relevant effect, not unstable**. Because t_0 is expressed by a sentence that involves a selfreferential attribution of truth, it is unstable with respect to t_0^T, and it is also plausible to assume that it is unstable* with respect to it (for one thing, $\mathfrak{C}$ does not entail that $t_0 \curvearrowright t_0 t_0^T$ holds). Yet, by ($\mathrm{A}_{\mathrm{P}}^{\curvearrowright}$) and (JUXT$^{\curvearrowright}$), $t_0 t_0 \curvearrowright t_0^T t_0^T$ holds, and so, by (CONTR$^{\curvearrowright}$), $t_0 \curvearrowright t_0^T t_0^T$ holds, and hence, if t_0 belongs to $\mathsf{ste}(\alpha)$, by the stability of $\uparrow$Snow is white$\uparrow$, t_0^T $\hat{}$ $\uparrow$Snow is white$\uparrow$ $\hat{}$ t_0^T—that is, $t_0 t_0^T$—belongs to $\mathsf{ste}(\alpha + 1)$. Therefore, t_0 causes t_0^T and does coobtain with it, but such coobtaining is partly due to whatever causes $\uparrow$Snow is white$\uparrow$ rather than wholly due to t_0 itself. t_0 is unstable* and yet not unstable**. A merely not unstable** SA *reoccurs* from the previous STE to the next STE, but, contrary to a stable one, does so without *sustaining itself* from the previous STE to the next STE.

Moreover, a similar case provides an example of a SA that is unstable and yet, since it causes its combination with the relevant effect, not unstable*. Let $t_1 = \uparrow\mathsf{exp}(t_1)$ is true$\uparrow$, so that $t_1 = t_1^T$. Because t_1 is expressed by a sentence that involves a selfreferential attribution of truth, it is unstable with respect to t_1^T. Yet, by ($\mathrm{A}_{\mathrm{P}}^{\curvearrowright}$) and (JUXT$^{\curvearrowright}$), $t_1 t_1 \curvearrowright t_1^T t_1^T$ holds, and so, by (CONTR$^{\curvearrowright}$), $t_1 \curvearrowright t_1^T t_1^T$—that is, $t_1 \curvearrowright t_1 t_1^T$—holds. Therefore, t_1 causes t_1^T and causes $t_1 t_1^T$, but what thus coobtains with the t_1^T-[token-effect] is not the t_1-[token-cause], but the t_1-[token-effect]. t_1 is unstable and yet not unstable* (see fn 47 for further discussion of t_1). A merely not unstable* SA *sustains itself* from the previous STE to the next STE, but, contrary to a stable one, does so without *continuing* from the previous STE to the next STE.

Back from the digression to our main thread, we can conclude that (CONTR$^{\curvearrowright}$) is unproblematic. By adding (DISTR$^{\curvearrowright}$), and so (CONTR$^{\curvearrowright}$), we do add that the causal role of a SA is at least as strong as that of its selfcombination, but, because of other features of causation (essentially, the absence of causal branching), that *apparent extra strength* turns out to be rather *nominal*: as we've seen in the eighth last paragraph, it basically boils down to causing not only certain SAs, but also their

selfcombinations. *The apparent extra strength is only measured by itself*, and so may not implausibly be taken to be actually null.[44]

This suggests that, more illuminatingly, $\mathfrak{C}$ could be reformulated by representing SAs as *sets* (or any other *extensional* kind of complex object) and combination as *union*, which would make more evident the vacuousness of selfcombination: similarly to how, given involutivity, complementation does not generate a new SA over and above s if the *complementandum* is s^*, so, given union idempotency, combination does not generate a new SA over and above s if the *combinanda* are s and s. It would not simply be that selfcombination does not generate a SA *with new (stronger) causal roles* (as (CONTR$^\curvearrowright$) already ensures); rather, selfcombination would not generate *a new SA* in the first place. Therefore, as far as causation is concerned, causes and effects could be taken to be composed in a very *extensional* manner: the identity of a combination is fully determined by *which elements compose it, independently of more specific facts of combination concerning them* (and so, as regards the particular issue in question, *independently of whether an element is selfcombined*). I'll henceforth assume this strong understanding of $\mathfrak{C}$ and also assume that an analogous treatment can be given to alternation,[45] so that, in particular, I'll henceforth assume that, in $\mathfrak{C}$, the composition operations are *idempotent*.

We can finally check that the previous *informal* considerations about the coherence of $\mathfrak{C}$ (*i.e.* {(POS$^\curvearrowright$), (EXH$^\curvearrowright$), (NEG$^\curvearrowright$), (EXC$^\curvearrowright$), (A$^\curvearrowright$), (STA$^\curvearrowright$), (JUXT$^\curvearrowright$), (DISTR$^\curvearrowright$)}) as a theory of nonnatural truth-related causation are not *formally* on the wrong track. To wit, $\mathfrak{C}$ is *consistent* and [*neutral with respect to* l] in the sense that:

Theorem 1 $\mathfrak{C}$ *does not entail that any chains* $\mathbf{p} \curvearrowright s \curvearrowright t \ldots \curvearrowright \mathbf{n}$, $l \curvearrowright s \curvearrowright t \ldots \curvearrowright \mathbf{n}$ *or* $\mathbf{p} \curvearrowright s \curvearrowright t \ldots \curvearrowright l$ *hold (where* $l = l^{T^*}$*).*

Proof Sketch. Observe first that any revision sequence $\mathcal{R}$ in the sense of [16] for the truth predicate T (letting $\mathfrak{T}$ be a suitable background theory including syntax, $\ulcorner\varphi\urcorner$

[44] Another possible worry about (CONTR$^\curvearrowright$) is that it entails that STEs to which, for example, l belongs are STEs to which ll—that is, ll^{T^*}—belongs; yet, ll^{T^*} is truth-relatedly impossible (a claim that will be vindicated in the further development of Section 5). However, even setting aside the fact that that's independently entailed by the extremely plausible clause given in the fifth last paragraph, that's simply a reflection of the fact that STEs are not *truth-relatedly possible situations*; rather, they are *elements in the causal structure of truth* (and are thus *prior* to truth-relatedly possible situations, helping to determine what is truth-relatedly possible or not in the way explained in Section 5). (Compare e.g. with the fact that, in the development of $\mathbb{N}$ of Section 3, at every stage, for some i, the arithmetically impossible SA $012\ldots i!$ obtains, see fn 20.) Unsurprisingly, in such structure, *divergence between* s *and* s^T *is the norm*: even for a vanilla SA like $\uparrow$Snow is white$\uparrow$ there is a STE to which it and $\uparrow$Snow is white$\uparrow^{T^*}$ both belong. While such divergence is *eventually eliminated* for *stable* SAs, it is *crucially ineliminable* for *unstable*** ones: reality and selfreferential ascending truth are doomed to be out of pace with one another (*pace* deflationism, see fn 32). Even in the further development of Section 5, if s is unstable**, we can have s but can't have it together with s^T; it is in the totalising nature of causal reality (Section 5) to determine that we then can have s together with s^{T^*}. Thanks to Sven Rosenkranz for comments that prompted this fn.

[45] For example, we could consider two set-like operations $\{\ldots\}_{\mathrm{C}}$ and $\{\ldots\}_{\mathrm{A}}$ to represent combination and alternation respectively. Setting the $\{\ldots\}_{\mathrm{C}}$-singleton ($\{\ldots\}_{\mathrm{A}}$-singleton) of a $\{\ldots\}_{\mathrm{A}}$-set ($\{\ldots\}_{\mathrm{C}}$-set), or of the representative of a complementation, or of the representative of a simple SA to be identical with its member, we could then let combination and alternation be represented by $\{\ldots\}_{\mathrm{C}}$-*union* and $\{\ldots\}_{\mathrm{A}}$-*union* respectively.

the numeral referring to the code of φ, $\|\varphi\|$ the set of sentences that are equivalent in $\mathfrak{T}$ with φ and $\mathsf{sen}(\|\varphi\|)$ a sentence in $\|\varphi\|$) is a model of $\mathfrak{C}$ which has:

- As domain, the set of nonempty sets of sentences that are equivalent in $\mathfrak{T}$;
- As interpretation of $\mathfrak{C}$'s singular terms, $\curvearrowright$ and 'STABLE' as it occurs in (STA$^{\curvearrowright}$), any function int such that:
  - If $\mathsf{int}(\tau) = \|\varphi\|$, $\mathsf{int}(\tau^*) = \|\neg\varphi\|$;
  - If $\mathsf{int}(\tau) = \|\varphi\|$ and $\mathsf{int}(\upsilon) = \|\psi\|$, $\mathsf{int}(\tau\upsilon) = \|\varphi \,\&\, \psi\|$;
  - If $\mathsf{int}(\tau) = \|\varphi\|$ and $\mathsf{int}(\upsilon) = \|\psi\|$, $\mathsf{int}(\tau|\upsilon) = \|\varphi \vee \psi\|$;
  - $\mathsf{int}(\mathbf{p}) = \|\top\|$;
  - $\mathsf{int}(\mathbf{n}) = \|\bot\|$;
  - If $\mathsf{int}(\tau) = \|\varphi\|$, $\mathsf{int}(\tau^T) = \|T^{\ulcorner}\mathsf{sen}(\|\varphi\|)^{\urcorner}\|$;
  - $\mathsf{int}(\tau \curvearrowright \upsilon)$ = True iff, for every α, if every member of $\mathsf{int}(\tau)$ holds at stage $\mathsf{str}(\alpha)$ in $\mathcal{R}$, every member of $\mathsf{int}(\upsilon)$ holds at $\mathsf{str}(\alpha + 1)$;
  - $\mathsf{int}(\text{'STABLE'}) = \{x$: for every α, if every member of x holds at $\mathsf{str}(\alpha)$, every member of x holds at $\mathsf{str}(\alpha + 1)\}$.

(By way of illustration, suppose that $\mathsf{int}(s) = \|\varphi\|$ and that every member of $\|\varphi\|$ holds at $\mathsf{str}(\alpha)$. Then $\mathsf{sen}(\|\varphi\|)$ holds at $\mathsf{str}(\alpha)$, and so, by the properties of $\mathcal{R}$, $T^{\ulcorner}\mathsf{sen}(\|\varphi\|)^{\urcorner}$ holds at $\mathsf{str}(\alpha + 1)$, and hence every member of $\|T^{\ulcorner}\mathsf{sen}(\|\varphi\|)^{\urcorner}\|$ holds at $\mathsf{str}(\alpha+1)$. Since $\mathsf{int}(s^T) = \|T^{\ulcorner}\mathsf{sen}(\|\varphi\|)^{\urcorner}\|$, that means that $s \curvearrowright s^T$ holds.) Recall then that, for some λ, λ is equivalent in $\mathfrak{T}$ with $\neg T^{\ulcorner}\lambda^{\urcorner}$. Letting $\mathsf{int}(\text{'}l\text{'}) = \|\lambda\|$ and $\mathsf{sen}(\|\lambda\|) = \lambda$ (so that $l = l^{T^*}$), notice finally that in no such model any chains $\mathbf{p} \curvearrowright s \curvearrowright t \ldots \curvearrowright \mathbf{n}$, $l \curvearrowright s \curvearrowright t \ldots \curvearrowright \mathbf{n}$ or $\mathbf{p} \curvearrowright s \curvearrowright t \ldots \curvearrowright l$ hold. □

Clearly, Theorem 1 can be extended to cover all other usual paradoxical sentences (like, for example, Curry sentences) and show that $\mathfrak{C}$ is neutral with respect to the SAs they express.[46]

[46] It should be clear that, not only the proof of Theorem 1, but large swathes of the picture behind the last two sections are heavily indebted to the tradition of *revision theory* (in particular, I drew much inspiration from [20]). (In fact, one could give a *foundational role to revision sequences*, so that the [revision-sequence]-based semantics of $\mathfrak{C}$ presented in the proof of Theorem 1 would ground—as its intended semantics—$\mathfrak{C}$, which in turn grounds—as its basic structure—the theory of truth-related implication in Section 5, which in turn grounds—as its intended semantics—our target logic, with the result that *it would be revision sequences that ultimately ground our target logic*. While discussion of this philosophical outlook (including its less fine-grained individuation of SAs) lies beyond the scope of this paper, I'll simply put on the record that I myself would rather give a *foundational role to* $\mathfrak{C}$ *itself* also with respect to revision sequences.) I agree with that tradition that revision sequences capture important aspects of the *metaphysical behaviour of truth*—in particular, the fact that its exemplification is *dynamic* in that, rather than *being reducible to a simple "yes-or-no"- (or "neither"-, or "both"-, or what have you) issue, it essentially evolves through STEs*, with sentences changing from truth to untruth and *vice versa*. As will become clear in Sections 5 and 6, where I do depart from the revision-theoretic tradition is *in extracting the logic out of such behaviour*: while that tradition typically extracts a *fairly classical* logic by focussing on what is *stable* in revision sequences, I prefer to extract instead a *fairly nonclassical* logic by focussing rather on what is *unstable* in them (being thereby inspired by the visionary *dictum* of [14], p. 5: "the process of revision can be performed by means of logical consequence"). In this connection, I should mention [35] as another recent proposal—alternative to the one I'm developing—relating revision theory and contraction. Standefer develops a logic that tracks revision sequences by *indexing* premises and conclusions (where φ^i is taken to express, roughly, that φ holds at the ith STE), for then observing that φ^i, φ^j cannot be

5 Implication

Let's now think about another prominent "arrow-ish" relation SAs can enter into: *implication*. (Notationally, just as I use $\curvearrowright$ to express *nonnatural truth-related causation*, I use $\rightarrow$ to express *truth-related implication*; terminologically, just as I use '*cause*' and '*effect*' to refer to the two arguments of causation, I use '*condition*' and '*consequence*' to refer to the two arguments of implication.) $\mathfrak{C}$ may be a good theory of nonnatural truth-related *causation*, but it is extremely problematic as a theory of truth-related *implication*. Immediately, the problem emerges because of the *weakness* of $\mathfrak{C}$. More concretely, the problem emerges because it is uncontroversial that *not every case of implication is a case of causation*, even when restricting to those of the truth-related variety (for example, as has already been observed in Section 4, it is uncontroversial that $\uparrow$Snow is white$\uparrow^T \rightarrow \uparrow$Snow is white$\uparrow$ holds but $\uparrow$Snow is white$\uparrow^T \curvearrowright \uparrow$Snow is white$\uparrow$ does not), and, in $\mathfrak{C}$, $\curvearrowright$ reflects that. More abstractly, the problem emerges because, in $\mathfrak{C}$, $\curvearrowright$ *lacks all the Tarski-closure properties*—that is, in addition to (MON), the causation principle of *reflexivity*:

(REFL) $\quad s \rightsquigarrow s$ holds[47]

"contracted" in the logic (to either φ^i or φ^j, I take it). However, since φ^i and φ^j are *not the same sentence* (*nor equivalent sentences*), that observation in itself would not seem to show that contraction fails in any interesting sense—so much so that, as Standefer himself notes (p. 69), the natural version of contraction in the logic (from φ^i, φ^i to φ^i) does hold in it. (Compare: in classical logic, Fx, Fy cannot be "contracted", but that hardly shows that contraction fails in classical logic in any interesting sense.) Contraction would indeed fail if the *indexed* logic (pp. 65–66) were used to define an *unindexed* logic by, a bit roughly, setting $\varphi_0, \varphi_1, \varphi_2 \ldots \vdash \psi_0, \psi_1, \psi_2 \ldots$ to hold in the unindexed logic iff, for some $i_0, i_1, i_2 \ldots$, $j_0, j_1, j_2 \ldots$, $\varphi_0^{i_0}, \varphi_1^{i_1}, \varphi_2^{i_2} \ldots \vdash \psi_0^{j_0}, \psi_1^{j_1}, \psi_2^{j_2} \ldots$ holds in the indexed logic—a move that Standefer himself does not make (p. 69 fn 31). However, such unindexed logic is rather unappealing: on the *undergeneration* side, because, as Standefer himself in effect notes (pp. 69–70), it invalidates the rightful metarules of the usual 2ary operators and intersubstitutability of φ with $T^{\ulcorner}\varphi^{\urcorner}$; on the *overgeneration* side, because it validates the contraction-compulsive $\oslash \vdash \varphi \rightarrow (\varphi \,\&\, \varphi)$ and, while it validates $\varphi \vdash T^{\ulcorner}\varphi^{\urcorner}$ and $T^{\ulcorner}\varphi^{\urcorner} \vdash \varphi$, it also validates the truth-repulsive $\oslash \vdash (\lambda \,\&\, \neg T^{\ulcorner}\lambda^{\urcorner}) \vee (\neg\lambda \,\&\, T^{\ulcorner}\lambda^{\urcorner})$. Having noted all this, I regard the broad idea of "*dropping the indices*" from the indexed logic as insightful, and, in fact, keeping in mind that indices represent STEs, that is what I myself will do in Section 5 by lifting cases of causation (in particular, (A)) to cases of implication. Indeed, an index-dropping policy different from the one criticised above would be to define the logic as the closure under the principles of the indexed logic *once indices are dropped from those*. Such new unindexed logic validates all the principles of our target logic; unfortunately, it also validates contraction (validating in effect the whole of classical logic *plus* naive truth), and is therefore trivial. The problem is that the indexed logic, even together with its official revision-theoretic semantics, does not offer any obvious reason for restoring nontriviality *by denying contraction (and so getting to our target logic) rather than by denying any other suitable combination of principles* of the new unindexed logic. It is in the more discriminating framework of $\mathfrak{C}$ (with its accompanying notions of instability and of a STE) that *the peculiar status of* (DISTR$^{\curvearrowright}$) *(and so of* (CONTR$^{\curvearrowright}$)*) emerges*, and that, consequently (as I'll substantiate in Section 5), *a reason becomes available for denying contraction rather than any other suitable combination of principles* of the new unindexed logic. Thanks to two anonymous referees for encouraging a development of the material in this fn.

[47] Those immersed in the contemporary literature on causation might be tempted to think that $\curvearrowright$ has at least one of the opposite properties, namely *irreflexivity*. But it's easy to see that, *given selfreferential ascending truth, $\curvearrowright$ is not irreflexive either*. By (A$_{\mathrm{P}}^{\curvearrowright}$), $t_1 \curvearrowright t_1^T$ holds, and so, since $t_1 = t_1^T$, $t_1 \curvearrowright t_1$ holds. Once truth is taken into account, causation is not irreflexive. It might be replied that that falls short of showing that a SA, t_1, *obtains* which is its own cause, and that, in fact, that is just one more reason for thinking that t_1 does *not* obtain. However, while the first part of the reply is certainly correct, the second

and the metacausation principle of *transitivity*:

(TRANS) If $s \rightsquigarrow t$ and $t \rightsquigarrow u$ hold, $s \rightsquigarrow u$ holds.[48]

But, clearly, on many plausible conceptions of implication, there are strong reasons for thinking that it has *all* the three Tarski-closure properties (for example, on the conception of implication as *necessary truth preservation*, or, better phrased in our context, as *necessary preservation of facts*, a conception that I assume).[49]

On the one hand, plausibly, *nonnatural causation constitutes the basic structure on which implication develops*, so that *all the core cases of nonnatural causation are cases of implication*: a theory of nonnatural causation thus provides at least the *beginnings* from which to extract a theory of implication. On the other hand, as I've variously mentioned in the last paragraph, implication is a relation *reaching beyond* nonnatural causation, so that, while $\mathfrak{C}$ does provide the beginnings from which to extract a theory of truth-related implication $\mathfrak{I}$, $\mathfrak{I}$ cannot be taken simply to *coincide* with $\mathfrak{C}$; rather, in developing $\mathfrak{I}$, $\mathfrak{C}$ needs to be *extended* with the Tarski-closure properties (which, nicely enough, will actually suffice to cover all cases of truth-related implication), and then *revised* in those respects in which it relied on the failure of those properties for nonnatural truth-related causation (*cf* fn 34). In developing $\mathfrak{I}$, we are thus led to add (REFL$^{\rightarrow}$), (MON$^{\rightarrow}$) and (TRANS$^{\rightarrow}$) to (POS$^{\rightarrow}$), (EXH$^{\rightarrow}$), (NEG$^{\rightarrow}$), (EXC$^{\rightarrow}$), (A$^{\rightarrow}$), (STA$^{\rightarrow}$) and (JUXT$^{\rightarrow}$) (*i.e.* to the totality of the implication and metaimplication principles corresponding to the causation and metacausation principles of $\mathfrak{C}$ *minus* (DISTR$^{\rightarrow}$), precisely because the justification for (DISTR$^{\curvearrowright}$) relied on the failure of the Tarski-closure properties for nonnatural truth-related causation, so that (DISTR$^{\rightarrow}$) will be criticised and replaced by another principle in the seventh next paragraph).

But, when we do so, a new, crucial issue emerges. We've seen in Section 4 the reasons why (CONTR$^{\curvearrowright}$) is unproblematic in $\mathfrak{C}$: these crucially included the fact that, in $\mathfrak{C}$, essentially because of the failure of the Tarski-closure properties, *unstable causes do not typically have intrastage effects* (so that causal branching does not occur).

---

part shoots itself in the foot. For, if t_1 does not obtain, t_1^* does, but also, by (A$_{\mathrm{N}}^{\curvearrowright}$), $t_1^* \curvearrowright t_1^{T^*}$ holds, and so, since $t_1 = t_1^T$, $t_1^* \curvearrowright t_1^*$ holds. Again, once truth is taken into account, causation is not irreflexive and, *either way, a SA obtains which is its own cause*. Talking about reflexivity, given the broad picture underlying this paper it should no longer come as a surprise that one of the fathers of Western logic was very much prone to theorising about implication in terms of causation, and that, by doing so, he pretty much committed himself to the irreflexivity of implication (as *per* the epigraph of this paper). Even setting aside the point that one should presumably distinguish between implication and causation, as he is also one of the fathers of Western theory of truth including (A) (*Metaphysics*, Γ, 7, 1011b), and aware to some extent of the unsettling effects of selfreferential ascending truth (*Sophistical Refutations*, 25, 180a–b), he should have known better ("*magis amica veritas*"?!). Moreover, and perhaps even more incisively for him personally, a promiscuous *Barbara* where major, middle and minor term are the same straightforwardly disproves irreflexivity of implication (*modulo* (CONTR$^{\rightarrow}$)). Indeed, since he accepted that $s \rightarrow s^T$ and $s^T \rightarrow s$ hold (*Categories*, 12, 14b), by (TRANS$^{\rightarrow}$) below in the text he should have accepted (REFL$^{\rightarrow}$) (*cf* fn 60). Thanks to an anonymous referee for suggesting a connection between this fn and fn 60.

[48] Notice that, together with (REFL) and (JUXT), (TRANS) implies the principles that, if $s \rightsquigarrow t$ and $tu \rightsquigarrow v$ hold, $su \rightsquigarrow v$ holds and that, if $s \rightsquigarrow t|u$ and $t \rightsquigarrow v$ hold, $s \rightsquigarrow v|u$ holds.

[49] Notice that, since, in our framework, implication relates *single* conditions with *single* consequences (*cf* fn 31), such conception does not clash with failure of contraction in the way indicated by [53].

But *the analogue of that fact for conditions and consequences does not hold in* $\mathfrak{I}$. By (REFL$^{\rightarrow}$), $l \rightarrow l$ holds, and that is an intrastage consequence of l; by (MON$^{\rightarrow}$), since $l \rightarrow l^T$ holds, $l \rightarrow l^T|l$—that is, $l \rightarrow l|l^*$—holds, and that is an (admittedly less exciting) intrastage consequence of l; by (TRANS$^{\rightarrow}$), since $l \rightarrow l^T$ and $l^T \rightarrow l$ hold,[50] $l \rightarrow l$ holds, and that is an intrastage consequence of l. Contrary to causation, implication has the *power*, given a condition, *either to develop it interstage or to keep it intrastage*. More generally, while causation has a more *material*, *concrete* character that, given any s, only links it with what s is immediately connected to by the causal joints along which reality articulates itself, implication has a more *idealising*, *abstracting* character that, given any s, links it with any element s combinationally contains or is alternationally contained in (thus yielding (MON$^{\rightarrow}$) and (REFL$^{\rightarrow}$)) and with any point of any causal chain starting with s (thus yielding (TRANS$^{\rightarrow}$) and, again, (REFL$^{\rightarrow}$)). Pictorially, implication can explore to any arbitrary depth each SA and to any arbitrary length each causal chain of SAs. By being Tarski-closed, implication is a reflection on causation—a light penetrating its dark structure.

Because unstable conditions do have intrastage consequences, contrary to causal branching *implicational branching* (*i.e.* a situation where a SA b_0 implies not only a SA b_1 with respect to which it is unstable, but also a SA b_2 that, because of that instability, does not coobtain with b_1,[51] with the effect that b_0 does not imply $b_1 b_2$) does occur. For example, $l \rightarrow l$ and $l \rightarrow l^T$ hold, but $l \rightarrow ll^T$ does not. To go back to a theme that emerged in Section 4, it is natural to put this more informally by saying that, although l implies l and implies l^T, and although it thus implies *either*, it does not imply *both*. Implication does have the power, given a condition, *either* to develop it interstage *or* to keep it intrastage, but it does not have the power to do *both*. Given what I've said in the last paragraph about the idealising, abstracting character of implication, I submit that such "either-but-not-both"-pattern is much more plausible for implication than for causation. For, while a relation cannot plausibly *correspond* in different directions to the causal joints along which reality articulates itself (since reality does not plausibly articulate itself in different directions in the first place!), a relation can plausibly *explore* in different directions elements of SAs and causal chains of SAs. And, while it is conceivable that the consequences reached by idealising and abstracting in different directions can also be reached together, it is also conceivable that, because of some sort of *Unschärfe* in the subject matter, reaching one consequence lying in one direction *precludes* reaching another consequence lying in another direction. In this sense, and as my use of the notion of power should already have prefigured, the modality with which a condition *leads* to a consequence,

[50] That $l^T \rightarrow l$ hold is an instance of ($D_P^{\rightarrow}$), and, as observed in Section 4, that principle follows from ($A_N^{\rightarrow}$) by (CONTRAP$^{\rightarrow}$). To see that, in turn, (CONTRAP$^{\rightarrow}$) holds, suppose that $s \rightarrow t$ holds. By (REFL$^{\rightarrow}$), $t^* \rightarrow t^*$ holds, and so, by (POS$_{\geq}^{\rightarrow}$) and (EXH$_{\geq}^{\rightarrow}$), $t^* \rightarrow t^*(s|s^*)$ holds, and hence, by (SEL$_{\restriction \leq}^{\rightarrow}$) below in the text, $t^* \rightarrow t^*s|s^*$ holds. Since $s \rightarrow t$ holds, by (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$) $t^*s \rightarrow t^*t$ holds, and so, by (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$), $t^*s|s^* \rightarrow t^*t|s^*$ holds, and hence, by (EXC$_{\leq}^{\rightarrow}$) and (NEG$_{\leq}^{\rightarrow}$), $t^*s|s^* \rightarrow s^*$ holds, and thus, by (TRANS$^{\rightarrow}$), $t^* \rightarrow s^*$ holds.

[51] Henceforth, I'll no longer understand as implicit the proviso 'if the obtaining of s token-causes the obtaining of t' (fn 15), and I'll implicitly rely instead on the fact that, for the kind of causation focussed on in this paper, 'If the obtaining of s token-causes the obtaining of t, s does not coobtain with t' plausibly entails 's does not coobtain with t'.

contrary to the modality with which a cause *leads* to an effect, is more akin to *the realisation of a possibility* than to *the enforcement of a necessity*. While causation is the realm of *law*, implication is the realm of *freedom*.

Because of this character of implication, in $\mathfrak{I}$ (JUXT$^{\rightarrow}$) does imply that the selfcombination of an unstable SA implies a combination of SAs that do not coobtain because of the instability of one with respect to the other (for example, $ll \rightarrow ll^T$ holds). Therefore, in $\mathfrak{I}$, the extremely plausible interpretation forced by (JUXT$^{\rightarrow}$) of the implicational strength of a combination does make the selfcombination of an unstable SA (not only *in principle* but) *as a matter of fact* implicationally much stronger than the SA itself. In $\mathfrak{I}$, given (JUXT$^{\rightarrow}$), the selfcombination of an unstable SA s has the power *both* to keep s intrastage *and* to develop it interstage, and so the power to *put together two SAs that, because of the instability of one with respect to the other, belong to different STEs and do not coobtain*. Such selfcombination thus in effect denies the instability of s. (For example, ll has the power *both* to keep l intrastage (as l) *and* to develop it interstage (as l^T), and so the power to *put together two SAs (l and l^T) that, because of the instability of one with respect to the other, belong to different STEs and do not coobtain*. ll thus in effect denies the instability of l.) And, since s still does not do any such thing, (CONTR$^{\rightarrow}$) does not hold.

Since this is supposed to be the main point of this paper, let's put it in a more canonical fashion. Suppose that s is unstable with respect to t in that:

(INSTA$_0$) $s \curvearrowright t$ holds;
(INSTA$_1$) st does not obtain.[52]

By (INSTA$_0$), $s \curvearrowright t$ holds, and so, typically, given the relation between causation and implication explained in the fourth last paragraph, $s \rightarrow t$ holds, and hence, by (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$), $ss \rightarrow st$ holds. Therefore, since, by (INSTA$_1$), st does not obtain, given that implication is necessary preservation of facts neither does ss. Now, if (CONTR$^{\rightarrow}_{\nsim}$) held, by (REFL$^{\rightarrow}$) so would $s \rightarrow ss$, and, given that ss does not obtain and that implication is necessary preservation of facts, that would imply that s does not obtain, which it might well do (instability does not *prevent* obtaining, as is indicated e.g. by the fact that l and l^* are unstable and yet, essentially by (POS$^{\rightarrow}_{\geq}$) and (EXH$^{\rightarrow}_{\geq}$), they contraobtain). Therefore, (CONTR$^{\rightarrow}_{\nsim}$) does not hold because it licenses the implication from a SA that might obtain to one that [does not because of instability]. The selfcombination of an unstable SA is surprisingly implicationally strong *vis-à-vis* the SA itself in that, given (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$), *it constitutes nothing less than a denial of the instability of the SA* (in particular, of (INSTA$_1$)).

A dual point holds for selfalternations. By (INSTA$_0$), $s \curvearrowright t$ holds, and so, typically, given the relation between causation and implication explained in the fifth last paragraph, $s \rightarrow t$ holds, and hence, by (CONTRAP$^{\rightarrow}$) (fn 50), $t^* \rightarrow s^*$ holds, and thus, by (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$), $s^*|t^* \rightarrow s^*|s^*$ holds. Therefore, since, by (INSTA$_1$), $s^*|t^*$ obtains,[53] given that implication is necessary preservation of facts

[52] Notice that (INSTA$_1$) plausibly follows from s's instability** with respect to t (fn 51) and that, in the development of this section, focus on instability** comes at no real loss of generality (fn 59).

[53] By (REFL$^{\rightarrow}$), (POS$^{\rightarrow}_{\geq}$) and (EXH$^{\rightarrow}_{\geq}$), $(st)^* \rightarrow (st)^*(s|s^*)$ holds, and so, by (SEL$^{\rightarrow}_{|\leq}$) below in the text, $(st)^* \rightarrow (st)^*s|s^*$ holds. Therefore, by (POS$^{\rightarrow}_{\geq}$) and (EXH$^{\rightarrow}_{\geq}$), $(st)^* \rightarrow ((st)^*s|s^*)(t|t^*)$ holds, and so, by

so does $s^*|s^*$. Now, if (CONTR$^{\rightarrow}_{|}$) held, by (REFL$^{\rightarrow}$) so would $s^*|s^* \rightarrow s^*$, and, given that $s^*|s^*$ obtains and that implication is necessary preservation of facts, that would imply that s^* obtains, which it might well not, since s might well do. Therefore, (CONTR$^{\rightarrow}_{|}$) does not hold because it licenses the implication from a SA that obtains because of instability to one that might not. The selfalternation of the complementation of an unstable SA is surprisingly implicationally weak *vis-à-vis* the complementation itself of the SA in that, given (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$), *it constitutes nothing more than an assertion of the instability of the SA* (in particular, of (an implicational equivalent of) (INSTA$_1$)).[54]

A notable consequence of the weakness of certain selfalternations is that, if one knows (proves, believes, supposes *etc.*) that a selfalternation $s|s$ obtains, and so if one knows that, *either way*, s obtains, it does not *follow* that one *knows* that s obtains (although it does follow that one *has a good reason for accepting* that s obtains, which in turn *nondeductively supports* the claim that one does know that s obtains, see (Zardini, E., Unstable knowledge, unpublished)). For, if s^* is unstable, that directly implies that $s|s$ obtains (as *per* the last paragraph), but s might well not—it might well be s^* that, in addition to being unstable, obtains instead. One knows a disjunction without knowing either disjunct not because of the *subjective* fact that one is in a somehow less than optimal epistemic position for deciding between two SAs characterising respectively the two possibilities opened by the disjunction; quite the contrary, one is past that stage of ignorance, since one knows of a single SA that characterises both possibilities. Rather, one knows a disjunction without knowing either disjunct because of the *objective* fact that the SA that one knows characterises both possibilities *is not forced* by those being the two possibilities opened by the disjunction (this can also be seen by reflecting that, given the instability of the selfalternated SA, one knows that both possibilities are also characterised by any of the effects with respect to which the SA is unstable, so that one might equally well infer to some such effect *instead*, see [48], p. 470 fn 16). *The disjunction opens the two possibilities* in a *radical*, *metaphysical* way rather than in a *superficial*, *epistemic* one: the fact that these are the two possibilities opened by the disjunction, even if they are characterised by a single SA, does not imply that they are resolved in reality in the obtaining of that SA. The two possibilities can remain open and reality itself unresolved. In other words, it is not that the relevant portion of reality has decided between possibility x characterised by s and possibility y characterised by t while one is still undecided about which of those two SAs characterises reality; rather, one has decided

(SEL$^{\rightarrow}_{\hat{|}\leq}$), $(st)^* \rightarrow ((st)^*s|s^*)t|t^*$ holds. Since, by (REFL$^{\rightarrow}$) and (SEL$^{\rightarrow}_{\hat{|}\leq}$), $((st)^*s|s^*)t \rightarrow (st)^*st|s^*$ holds, by (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$) $((st)^*s|s^*)t|t^* \rightarrow (st)^*st|s^*|t^*$ holds, and so, by (TRANS$^{\rightarrow}$), $(st)^* \rightarrow (st)^*st|s^*|t^*$ holds, and hence, by (EXC$^{\rightarrow}_{\leq}$) and (NEG$^{\rightarrow}_{\leq}$), $(st)^* \rightarrow s^*|t^*$ holds.

[54] In the specific case of a SA like l, the strength of selfcombination and weakness of selfalternation can be made to emerge also in a different, more straightforward way by replacing appeal to *instability* with appeal to *the properties of truth*. $ll = ll^{T^*}$ and $l^*l^* = l^*l^T$, but the latter are truth-theoretic barbarities; $l|l = l|l^{T^*}$ and $l^*|l^* = l^*|l^T$, but the latter are truth-theoretic matters of course. (Recall however that, sometimes (fn 44), barbarities are committed and matters of course omitted.) Short of a weird behaviour of complementation, failure of (CONTR$^{\rightarrow}$) is virtually mandated by the properties of truth [17, 44].

that s characterises both x and y, but the relevant portion of reality is still undecided between x and y.[55] Instability fractures reality into unresolvable possibilities.[56]

Since (CONTR$^{\rightarrow}$) does not hold, neither does (DISTR$^{\rightarrow}$), whereas, by (REFL$^{\rightarrow}$) and (JUXT$^{\rightarrow}$), (SEP$^{\rightarrow}$) holds (as befits the abstracting power of implication). In the problematic case $s(t|u)$ where st is possible in $\mathfrak{C}$ but su is impossible, (DISTR$^{\rightarrow}$) would give implication the power both to develop s interstage (in order to try to make st impossible) and to keep it intrastage (in order to make su impossible). But, as we've already stressed, that is a power that implication does not have: implication only has the power either to develop s interstage or to keep it intrastage. Generalising from the features emerged in this discussion, instead of (DISTR), $\mathfrak{I}$ only contains the metaimplication principle of *selection*:

(SEL$^{\wedge}_{|\leq}$) If $s \rightsquigarrow t(u|v)$ holds, $s \rightsquigarrow tu|v$ holds, and, if $st|u \rightsquigarrow v$ holds, $s(t|u) \rightsquigarrow v$ holds.[57]

Notice the great difference between (DISTR$^{\wedge}_{|\leq}$) and (SEL$^{\wedge}_{|\leq}$). (DISTR$^{\wedge}_{|\leq}$) allows us to go from a combination of a *local combinandum* with an alternation of *local alternanda* to an alternation of *global* combinations.[58] (SEL$^{\wedge}_{|\leq}$) does not do that; it only allows us to go from a combination of a *local combinandum* with an alternation of *local alternanda* to an alternation of a *global* combination with a *local alternandum*. Therefore, contrary to (DISTR$^{\wedge}_{|\leq}$), in general, (SEL$^{\wedge}_{|\leq}$) complies with the maxim "Stay local", and, in particular, does not imply that a possible SA can be expanded, by repeated applications of (POS$_{\geq}$) and (EXH$_{\geq}$), into an alternation of

[55] By way of anticlimax, I think that, among the (possibly few!) merits of this view, there's at least the one of vindicating Yogi Berra's memorable *dictum* "When you come to a fork in the road, take it" ([3], p. 9), which I understand as recommending to stick to the fork rather than going for one of the roads leading from the fork *even if those roads lead to the same place and even if such place lies right at the start of the roads* (the punch coming obviously from the italicised clause). I thereby disagree with another commentator of Berra ([11], pp. 172–173), who rather understands the *dictum* as literally making the opposite claim that one should go for either road (which would seem to have been Berra's own understanding of his *dictum* as restricted to the case where the roads lead to the same place—an understanding that, so restricted, while it does not apply at the level of *implication*, as I've already partially indicated in the text does apply at the level of *nondeductive support*), and, noting that that is not always sensible, proposes a charitable reinterpretation of the *dictum* (let's call it 'the Field-Berra *dictum*') to the effect that "if you come to a fork in the road, and know that neither of the two roads leading from the fork will take you to your desired destination, go back and try a different route". For what it's worth, both $\mathfrak{C}$ and $\mathfrak{I}$ vindicate the Field-Berra *dictum* (since, if $s \rightsquigarrow \mathbf{n}|\mathbf{n}$ holds, so does $s \rightsquigarrow \mathbf{n}$). (Notice however that $\mathfrak{C}$, contrary to $\mathfrak{I}$, violates my understanding of Berra's *dictum*, since it validates (CONTR$^{\curvearrowright}_{|}$). Unresolvable possibilities only arise at the level of implication, thus corresponding to its characteristic freedom.) Also, I agree with Field that supervaluationist theories violate Berra's *dictum*, but not so much because they violate the Field-Berra *dictum* (which they certainly do), but because they violate my understanding of Berra's *dictum* (as does virtually every other contractive theory as well as noncontractive theories with an additive treatment of disjunction as opposed to the multiplicative one offered in this paper).

[56] Thanks to Sven Rosenkranz for pressing me on the issues discussed in this paragraph.

[57] $\mathfrak{I}$ does not contain the analogues for selection of (DISTR$^{\wedge}_{|\geq}$) and (DISTR$_{|^{\wedge}\leq}$) because they are not even classically valid. Meanwhile, the analogue for selection of (DISTR$_{|^{\wedge}\geq}$) is just a notational variant of (SEL$^{\wedge}_{|\leq}$).

[58] Throughout, I understand the local/global distinction so that, while s, t and their relatives count as local, st and its relatives count as global. The choice of terminology should become clearer by the end of this paragraph.

maximal SAs (*i.e.* combinations such that, for every relevant s, either s is one of their *combinanda* or s^* is)—an alternation that would be impossible in $\mathfrak{I}$, given that, as witnessed e.g. by the *Postcard* paradox, there are SAs that cannot jointly be supposed either to obtain or not to obtain on pain of contradicting $(A^{\rightarrow})$ (see [50]; (Zardini, E., Against the world, unpublished) for more discussion on this issue). The light of implication only projects partial likenesses of what was the totality of causal reality.

Let's close this section by putting what has been done in a broader perspective. What instability of s with respect to t most directly requires, from causation as well as from implication, is failure of the principle of *persistence*:

(PERS) If $s \rightsquigarrow t$ holds, $s \rightsquigarrow st$ holds, and, if $t \rightsquigarrow s$ holds, $s|t \rightsquigarrow s$ holds,

and, arguably, the semantic paradoxes are essentially solved as soon as one gives up (PERS) (*cf* [47]). Therefore, contrary to what e.g. [42] seems to suggest, the principle whose failure at the *logical* level is most directly associated with *instability* at the *metaphysical* level *is not* $(CONTR^{\rightarrow})$, *is* $(PERS^{\rightarrow})$ (as witnessed, among other things, by the fact that $(PERS^{\curvearrowright})$ does not hold—it's a direct denial of instability*—whereas $(CONTR^{\curvearrowright})$ does). Insofar as one supports a solution to the semantic paradoxes by appealing to the idea of instability, *the most natural kind* that such solution instantiates is not the natural kind of denying $(CONTR^{\rightarrow})$, it is the natural kind of denying $(PERS^{\rightarrow})$—if the metaphysical root of paradox is instability, its logical root is $(PERS^{\rightarrow})$,[59] not $(CONTR^{\rightarrow})$. The fact is however that the problematic

[59] In general, when operating also at the level of implication, some claims that were the case when operating only at the level of causation are no longer the case. Without going into a full revisitation of causation by the lights of implication, let's at least see how the differences among instability, instability* and instability** look like from the point of view of implication. Say that a SA is *unfounded* iff neither its truth nor its untruth are implied by nonarbitrary facts (as is typically but not always the case for unstable SAs); *founded* otherwise. (In turn, by way of ostensive definition, if t_1 obtains it is an *arbitrary* fact while ↑Snow is white↑ and $t_1|t_1^*$ are *nonarbitrary* facts.) Then, since, when operating also at the level of implication, combination finally *manifests its full strength*, when operating also at the level of implication (PERS) fails *across the board* for unstable unfounded SAs. For example, when operating only at the level of causation, as *per* Section 4, t_1 was unstable but not unstable*, but that was merely because, given the strong understanding of $\mathfrak{C}$ assumed in Section 4, $t_1 t_1^T$ was identical with t_1. When operating only at the level of causation, instability of unfounded SAs was forced to assume a *token*-centred appearance because "selfcombinations" were *combinations only in a nominal sense* and therefore *did not represent a sense of coobtaining strong enough to capture the idea of instability*. Once, when operating also at the level of implication, combination manifests its full strength, selfcombinations do represent a sense of coobtaining strong enough for instability of unfounded SAs to be no longer forced to assume a token-centred appearance, so that we can now take t_1 and its relatives to be not only unstable, but also unstable*. Indeed, the same kind of consideration revolving around the fact that, when operating only at the level of causation, combination did not manifest its full strength, also shows that, when operating also at the level of implication, instability* of unfounded SAs is no longer forced to assume a *cause*-centred appearance, so that, wrapping up, we can now take every unstable unfounded SA to be not only unstable or unstable*, but also unstable**. (Notice that, while as far as *unfounded* SAs are concerned, instability's and instability*'s token- and cause-centredness respectively are thus an appearance that is overcome when operating also at the level of implication, as far as *founded* SAs are concerned instability's and instability*'s token- and cause-centredness respectively are a reality that is even more inescapable when operating also at the level of implication. To wit, letting $t_2 = \uparrow \mathsf{exp}(t_2)$ is true↑ | ↑$\mathsf{exp}(t_2)$ is not true↑, so that $t_2 = t_2^T|t_2^{T^*}$, essentially by $(POS^{\curvearrowright}_{\geq})$ and $(EXH^{\curvearrowright}_{\geq})$ $t_2 \curvearrowright t_2 t_2^T$ holds; letting $t_3 = \uparrow \mathsf{exp}(t_3)$ is true↑ | ↑Snow is white↑, so that $t_3 = t_3^T$ | ↑Snow is white↑, essentially by the stability of ↑Snow is white↑* and $(STA^{\rightarrow})$ $(t_3 t_3^T)^* \rightarrow$ ↑Snow is white↑* holds.)

instances of (PERS$^{\rightarrow}$) are implied by the triad consisting of (TRANS$^{\rightarrow}$), (JUXT$^{\rightarrow}$) and (CONTR$^{\rightarrow}$),[60] and so, if one wishes to deny (PERS$^{\rightarrow}$) effectively, one should deny one of these principles (well, one should at least deny the conjunction of the members of one nonempty subset of the set of these principles, but I take it that the nonsingleton options are deeply unsatisfactory if not for other reasons because they do not point to exactly what kind of logic is needed). Therefore, either one goes for some sort of *one-step* logic (revolving around failure of (TRANS$^{\rightarrow}$)), or one goes for some sort of *defeasible* logic (revolving around failure of (JUXT$^{\rightarrow}$)), or one goes for some sort of *nonidempotent* logic (revolving around failure of (CONTR$^{\rightarrow}$)).[61] I strongly favour, and have thus focussed on (here as elsewhere), the third option,[62] but it is now important to note that the present investigation of instability has been heuristically useful in uncovering other possible approaches to the logic of selfreferential ascending truth which can be supported by appealing to the idea of instability but which, to the best of my knowledge, have yet to be pursued. Let a thousand logics bloom.

6 Logical Consequence

It now remains to ascend from the *nonlinguistic* level of *SAs* at which *causation* and *implication* operate to the *linguistic* level of *sentences* at which *logical consequence* operates and show that $\mathfrak{I}$ (*i.e.* {(POS$^{\rightarrow}$), (EXH$^{\rightarrow}$), (NEG$^{\rightarrow}$), (EXC$^{\rightarrow}$), (A$^{\rightarrow}$), (STA$^{\rightarrow}$), (JUXT$^{\rightarrow}$), (SEL$^{\rightarrow}_{\hat{}\mid\leq}$), (REFL$^{\rightarrow}$), (MON$^{\rightarrow}$), (TRANS$^{\rightarrow}$)}) generates our target logic (*i.e.* the $\Rightarrow$*-free fragment* of [46]'s $\mathbf{IKT}_{\Rightarrow\mathsf{tf}}$, henceforth simply '**IKT**'). **IKT** can very naturally be presented in *sequent-calculus* format. It is the smallest logic containing as *axiom* the *structural rule*:

$$\frac{}{\varphi \vdash_{\mathbf{IKT}} \varphi}\ \mathrm{I}$$

and *closed* under the *structural metarules*:

$$\frac{\Gamma_0 \vdash_{\mathbf{IKT}} \Delta}{\Gamma_0, \Gamma_1 \vdash_{\mathbf{IKT}} \Delta}\ \text{K-L} \qquad \frac{\Gamma \vdash_{\mathbf{IKT}} \Delta_0}{\Gamma \vdash_{\mathbf{IKT}} \Delta_0, \Delta_1}\ \text{K-R}$$

$$\frac{\Gamma_0 \vdash_{\mathbf{IKT}} \Delta_0, \varphi \qquad \Gamma_1, \varphi \vdash_{\mathbf{IKT}} \Delta_1}{\Gamma_0, \Gamma_1 \vdash_{\mathbf{IKT}} \Delta_0, \Delta_1}\ \mathrm{S}$$

[60] Since $l \rightarrow l^T$ and $l^T \rightarrow l$ hold, by (TRANS$^{\rightarrow}$) $l \rightarrow l$ holds, and so, by (JUXT$^{\rightarrow}$), $ll \rightarrow ll^T$ holds, and hence, by (CONTR$^{\rightarrow}$), $l \rightarrow ll^T$ holds. Incidentally, the fact that $l \rightarrow l^T$, $l^T \rightarrow l$ and (TRANS$^{\rightarrow}$) together entail $l \rightarrow l$ shows that a *purely nonreflexive* approach to the semantic paradoxes as pursued e.g. by [12] (which accepts all the other basic principles of classical logic) is unviable (as long as one aims at validating basic principles of truth-related implication such as (A$^{\rightarrow}_{\mathrm{P}}$) and (D$^{\rightarrow}_{\mathrm{P}}$), *cf* fn 47).

[61] It is revealing that the leading examples used by [14] to motivate failure of contraction are immediately counterexamples to what is in effect (PERS) rather than (CONTR) (I suppose that Girard is implicitly assuming (REFL) (fn 60) and (JUXT)).

[62] That might be a bit misleading. In the case of nonnatural truth-related causation, I strongly favour, and have thus focussed on (in Section 4), a one-step theory, denying defeasible and nonidempotent theories.

(with S's being eliminable in **IKT**, see [46], p. 365) as well as under the *operational metarules*:

$$\frac{\Gamma \vdash_{\mathbf{IKT}} \Delta, \varphi}{\Gamma, \neg\varphi \vdash_{\mathbf{IKT}} \Delta}\ \neg\text{-L} \qquad \frac{\Gamma, \varphi \vdash_{\mathbf{IKT}} \Delta}{\Gamma \vdash_{\mathbf{IKT}} \Delta, \neg\varphi}\ \neg\text{-R}$$

$$\frac{\Gamma, \varphi, \psi \vdash_{\mathbf{IKT}} \Delta}{\Gamma, \varphi \,\&\, \psi \vdash_{\mathbf{IKT}} \Delta}\ \&\text{-L} \qquad \frac{\Gamma_0 \vdash_{\mathbf{IKT}} \Delta_0, \varphi \qquad \Gamma_1 \vdash_{\mathbf{IKT}} \Delta_1, \psi}{\Gamma_0, \Gamma_1 \vdash_{\mathbf{IKT}} \Delta_0, \Delta_1, \varphi \,\&\, \psi}\ \&\text{-R}$$

$$\frac{\Gamma_0, \varphi \vdash_{\mathbf{IKT}} \Delta_0 \qquad \Gamma_1, \psi \vdash_{\mathbf{IKT}} \Delta_1}{\Gamma_0, \Gamma_1, \varphi \vee \psi \vdash_{\mathbf{IKT}} \Delta_0, \Delta_1}\ \vee\text{-L} \qquad \frac{\Gamma \vdash_{\mathbf{IKT}} \Delta, \varphi, \psi}{\Gamma \vdash_{\mathbf{IKT}} \Delta, \varphi \vee \psi}\ \vee\text{-R}$$

$$\frac{\Gamma_0 \vdash_{\mathbf{IKT}} \Delta_0, \varphi \qquad \Gamma_1, \psi \vdash_{\mathbf{IKT}} \Delta_1}{\Gamma_0, \Gamma_1, \varphi \supset \psi \vdash_{\mathbf{IKT}} \Delta_0, \Delta_1}\ \supset\text{-L} \qquad \frac{\Gamma, \varphi \vdash_{\mathbf{IKT}} \Delta, \psi}{\Gamma \vdash_{\mathbf{IKT}} \Delta, \varphi \supset \psi}\ \supset\text{-R}$$

$$\frac{\Gamma, \varphi \vdash_{\mathbf{IKT}} \Delta}{\Gamma, T\ulcorner\varphi\urcorner \vdash_{\mathbf{IKT}} \Delta}\ T\text{-L} \qquad \frac{\Gamma \vdash_{\mathbf{IKT}} \Delta, \varphi}{\Gamma \vdash_{\mathbf{IKT}} \Delta, T\ulcorner\varphi\urcorner}\ T\text{-R}$$

$$\frac{\oslash \vdash_{\mathbf{IKT}} \Delta}{\mathsf{t} \vdash_{\mathbf{IKT}} \Delta}\ \text{t-L} \qquad \frac{}{\oslash \vdash_{\mathbf{IKT}} \mathsf{t}}\ \text{t-R}$$

$$\frac{}{\mathsf{f} \vdash_{\mathbf{IKT}} \oslash}\ \text{f-L} \qquad \frac{\Gamma \vdash_{\mathbf{IKT}} \oslash}{\Gamma \vdash_{\mathbf{IKT}} \mathsf{f}}\ \text{f-R}$$

(We also assume that **IKT** is expressive enough in terms of selfreference.)

Given that **IKT** is a *pure* logic that makes *no nonlogical assumptions about which SAs are stable*, for the purposes of this section we'll consequently understand 'STABLE' as it occurs in (STA$^{\rightarrow}$) as applying only to $\mathbf{p}$ and $\mathbf{n}$. Let then I be the set of mappings from finite multisets of sentences of the language of **IKT** to SAs such that, for every $\mathsf{int} \in I$:

- $\mathsf{int}(\neg\varphi) = \mathsf{int}(\varphi)^*$;
- $\mathsf{int}(\varphi \,\&\, \psi) = \mathsf{int}(\varphi)\hat{}\mathsf{int}(\psi)$;
- $\mathsf{int}(\varphi \vee \psi) = \mathsf{int}(\varphi)|\mathsf{int}(\psi)$;
- $\mathsf{int}(\varphi \supset \psi) = (\mathsf{int}(\varphi)\hat{}\mathsf{int}(\psi)^*)^*$;
- $\mathsf{int}(T\ulcorner\varphi\urcorner) = \mathsf{int}(\varphi)^T$;
- $\mathsf{int}(\mathsf{t}) = \mathbf{p}$;
- $\mathsf{int}(\mathsf{f}) = \mathbf{n}$,

where, for every multiset Γ, $\mathsf{int}_{\mathsf{P}}(\Gamma)$ combines the int-interpretations of the elements of Γ whereas $\mathsf{int}_{\mathsf{C}}(\Gamma)$ alternates them (if Γ is the empty multiset, let $\mathsf{int}_{\mathsf{P}}(\Gamma) = \mathbf{p}$ and

$\mathsf{int}_{\mathsf{C}}(\Gamma) = \mathbf{n}$). Say that $\Gamma \vdash_{\mathfrak{I}} \Delta$ holds iff, for every int, $\mathsf{int}_{\mathsf{P}}(\Gamma) \rightarrow \mathsf{int}_{\mathsf{C}}(\Delta)$ holds in $\mathfrak{I}$.[63] Then:

Theorem 2 *$\Gamma \vdash_{\mathfrak{I}} \Delta$ holds iff $\Gamma \vdash_{\mathbf{IKT}} \Delta$ holds.*

Proof Sketch.

- *Left to right.* Letting $T^{\ulcorner}\varphi^{\urcorner}$ count as nonatomic, say that int is *corresponding* iff int is a bijection between the set of atomic sentences and the set of simple SAs and is such that $\mathsf{int}(T^{\ulcorner}\varphi^{\urcorner}) = \mathsf{int}(T^{\ulcorner}\psi^{\urcorner})$ only if both $T^{\ulcorner}\varphi^{\urcorner} \vdash_{\mathbf{IKT}} T^{\ulcorner}\psi^{\urcorner}$ and $T^{\ulcorner}\psi^{\urcorner} \vdash_{\mathbf{IKT}} T^{\ulcorner}\varphi^{\urcorner}$ hold. Prove then, mainly by induction on the complexity of SAs, the lemma that, if int is corresponding, for every s, for some φ, $\mathsf{int}(\varphi) = s$, and, mainly by induction on the complexity of sentences, the lemma that, if int is corresponding, $\mathsf{int}(\varphi) = \mathsf{int}(\psi)$ only if both $\varphi \vdash_{\mathbf{IKT}} \psi$ and $\psi \vdash_{\mathbf{IKT}} \varphi$ hold. Now, say that $s \rightarrow_{\mathbf{IKT}} t$ holds iff, for every corresponding int, for every φ and ψ such that $s = \mathsf{int}(\varphi)$ and $t = \mathsf{int}(\psi)$, $\varphi \vdash_{\mathbf{IKT}} \psi$ holds. Observe then that $\rightarrow_{\mathbf{IKT}}$ validates all of (POS), (EXH), (NEG), (EXC), (A), (STA), (JUXT), ($\mathrm{SEL}^{\wedge}_{|\leq}$), (REFL), (MON) and (TRANS). (By way of illustration, suppose that $s \rightarrow_{\mathbf{IKT}} t(u|v)$ holds and that $\mathsf{int}(\varphi) = s$ and $\mathsf{int}(\psi) = tu|v$. Observe that it follows from the above lemmas and the second conjunct of the latter supposition that, for some χ_0, χ_1 and χ_2, $t = \mathsf{int}(\chi_0)$, $u = \mathsf{int}(\chi_1)$, $v = \mathsf{int}(\chi_2)$ and both $\psi \vdash_{\mathbf{IKT}} (\chi_0 \,\&\, \chi_1) \vee \chi_2$ and $(\chi_0 \,\&\, \chi_1) \vee \chi_2 \vdash_{\mathbf{IKT}} \psi$ hold. Now, by I, $\chi_0 \vdash_{\mathbf{IKT}} \chi_0$ and $\chi_1 \vdash_{\mathbf{IKT}} \chi_1$ hold, and so, by &-R, $\chi_0, \chi_1 \vdash_{\mathbf{IKT}} \chi_0 \,\&\, \chi_1$ holds. But, by I, $\chi_2 \vdash_{\mathbf{IKT}} \chi_2$ also holds, and so, by $\vee$-L, $\chi_0, \chi_1 \vee \chi_2 \vdash_{\mathbf{IKT}} (\chi_0 \,\&\, \chi_1) \vee \chi_2$ holds, and hence, by &-L, $\chi_0 \,\&\, (\chi_1 \vee \chi_2) \vdash_{\mathbf{IKT}} (\chi_0 \,\&\, \chi_1) \vee \chi_2$ holds. But, since, by supposition, $s \rightarrow_{\mathbf{IKT}} t(u|v)$ holds, $\varphi \vdash_{\mathbf{IKT}} \chi_0 \,\&\, (\chi_1 \vee \chi_2)$ holds, and so, by S, $\varphi \vdash_{\mathbf{IKT}} (\chi_0 \,\&\, \chi_1) \vee \chi_2$ holds, and hence, by the above observation and S, $\varphi \vdash_{\mathbf{IKT}} \psi$ holds. Since int, φ and ψ were arbitrary, that means that $s \rightarrow_{\mathbf{IKT}} tu|v$ holds.) Therefore, if $s \rightarrow t$ holds in $\mathfrak{I}$, $s \rightarrow_{\mathbf{IKT}} t$ holds. From this and the existence of a corresponding interpretation, infer finally that, if $\Gamma \vdash_{\mathfrak{I}} \Delta$ holds, $\Gamma \vdash_{\mathbf{IKT}} \Delta$ holds.
- *Right to left.* Observe first that $\vdash_{\mathfrak{I}}$ validates all the defining principles of **IKT**. (By way of illustration, suppose that $\Gamma, \varphi \vdash_{\mathfrak{I}} \Delta$ holds. Then, for every int, $\mathsf{int}_{\mathsf{P}}(\Gamma, \varphi) \rightarrow \mathsf{int}_{\mathsf{C}}(\Delta)$ holds (in $\mathfrak{I}$). Letting $\mathsf{int}_{\mathsf{P}}(\Gamma) = s$, $\mathsf{int}(\varphi) = t$ and $\mathsf{int}_{\mathsf{C}}(\Delta) = u$, that implies that $st \rightarrow u$ holds. Now, by ($\mathrm{REFL}^{\rightarrow}$), $t^* \rightarrow t^*$ holds, and so, by ($\mathrm{JUXT}^{\rightarrow}$), $st|t^* \rightarrow u|t^*$ holds, and hence, by ($\mathrm{SEL}^{\rightarrow}_{\wedge|\leq}$), $s(t|t^*) \rightarrow u|t^*$ holds, and thus, by ($\mathrm{EXH}^{\rightarrow}_{\geq}$) and ($\mathrm{POS}^{\rightarrow}_{\geq}$), $s \rightarrow u|t^*$ holds. Since

[63] Going back to issues mentioned in Section 2, notice that this style of definition implies that, keeping fixed the interpretation of the relevant expressions, 'Socrates is a better philosopher than Cicero' entails 'Socrates is a better philosopher than Tully' (since ↑Socrates is a better philosopher than Cicero↑ is the same SA as ↑Socrates is a better philosopher than Tully↑), but does not entail 'Socrates is a better philosopher than Cicero and tigers are animals' (since ↑Socrates is a better philosopher than Cicero↑ does not imply ↑Tigers are animals↑). More subtly (fn 11), notice also that this style of definition implies that, keeping fixed the interpretation of the relevant expressions, 'Socrates does not exist' entails '↑Socrates does not exist↑ does not obtain' (since ↑Socrates does not exist↑ causes ↑↑Socrates does not exist↑ does not exist↑), but does not entail 'Socrates exists' (since ↑↑Socrates does not exist↑ does not exist↑ does not imply ↑Socrates exists↑).

int was arbitrary, that means that $\Gamma \vdash_{\mathfrak{I}} \Delta, \neg\varphi$ holds.) From this, infer then that, if $\Gamma \vdash_{\mathbf{IKT}} \Delta$ holds, $\Gamma \vdash_{\mathfrak{I}} \Delta$ holds. □

Technically, and using our knowledge of **IKT** to illuminate $\mathfrak{I}$, this implies that (CONTR$^{\rightarrow}$) does not hold in $\mathfrak{I}$. *Philosophically*, and using our knowledge of $\mathfrak{I}$ to illuminate **IKT**, this explains why the metarules of *logical contraction*:

$$\frac{\Gamma, \varphi, \varphi \vdash \Delta}{\Gamma, \varphi \vdash \Delta}\ \text{W-L} \qquad \frac{\Gamma \vdash \Delta, \varphi, \varphi}{\Gamma \vdash \Delta, \varphi}\ \text{W-R}$$

do not hold in **IKT**. That is so because (CONTR$^{\rightarrow}$) does not hold in $\mathfrak{I}$, and, as we've seen in Section 5, that is in turn so because selfcombinations actually deny the instability of the selfcombined SA while selfalternations actually assert the instability of the complementation of the selfalternated SA.[64] Contraction fails because of the instability of certain SAs.[65]

Acknowledgments Earlier versions of the material in this paper have been presented in 2014 at the SILFS Workshop *Current Trends in the Philosophy of Logic* (University of Rome Three) and at the PERSP Metaphysics Seminar (University of Barcelona); in 2015, at the LanCog Seminar (University of

[64] We also get a *unified account of semantic paradoxicality*. Let $t_4 = (\uparrow \mathsf{exp}(t_4)\ \text{is true} \uparrow \hat{}\ \mathbf{p}^*)^*$, so that $t_4 = (t_4^T \mathbf{p}^*)^*$, and let $\kappa = \mathsf{exp}(t_4) = T^{\ulcorner}\kappa^{\urcorner} \supset \mathsf{t}$. *$\kappa$ is a plainly true sentence, and yet gives rise to a paradoxical, Curry-style derivation of* t (*cf* [24], p. 246; [25], pp. 472–473) which employs W-L for κ. We can account for κ's paradoxicality in terms of the instability of t_4, as that SA is expressed by a sentence—κ—which involves a selfreferential attribution of truth. Since t_4 is unstable (and indeed unstable* although not unstable**, see fn 59), (STA$^{\rightarrow}$) is not triggered for t_4, and so W-L for κ is not *built into* our system, even once this is extended with respect to **IKT** to include the stability of SAs other than **p** and **n**. It is true that W-L for κ is nevertheless *derivable in* **IKT** already, essentially using the facts that $\oslash \vdash_{\mathbf{IKT}} \mathsf{t}$ holds and that $\mathsf{t} \vdash_{\mathbf{IKT}} \kappa$ holds. But such derivation blatantly relies on an *antecedent proof* of t, and so the Curry-style derivation of t employing W-L for κ cannot plausibly be taken to offer a *new, independent proof* of t. Just as it is ultimately because of the instability of the SA expressed by a Curry sentence with an untrue consequent that Curry-style reasoning does not prove its consequent, it is because of the instability of the SA expressed by κ that Curry-style reasoning does not prove t. A similar point applies to any Curry sentence whose consequent is *true* and expresses a *stable* SA by considering the extension of our system with respect to **IKT** which includes such truth and stability. Notice that, while κ does have the property of expressing an unstable SA, *it lacks many of the properties that are usually taken to account for paradoxicality* (for example, there is no reasonable sense in which κ is *indeterminate*, as it is instead *plainly true*, *cf* [48], pp. 477, 482–483, 492 fn 45). While still on the topic of unification, I should further note that the instability approach to the semantic paradoxes can also be extended to the *epistemic paradoxes and related kinds of paradoxes* (I give the details in (Zardini, E., Unstable knowledge, unpublished)).

[65] I don't see that an *instability* approach to the semantic paradoxes *in particular* generates *revenge* issues, but a *noncontractive* approach *in general* might. Let $\lambda^{®}$ be the sentence 'It is not the case that $\lambda^{®}$ is true and contracts'. Then '$\lambda^{®}$ is true and contracts' entails $\lambda^{®}$ (*i.e.* its negation), and so it might be thought that, if '$\lambda^{®}$ is true and contracts' contracts, an untruth eventually follows. By contraposition on logical consequence, '$\lambda^{®}$ is true and contracts' does not contract, and so, since, in general, φ contracts iff $\neg\varphi$ does, its negation—that is, $\lambda^{®}$—does not contract either, and hence, *a fortiori*, it is not the case that $\lambda^{®}$ is true and contracts, which is tantamount to $\lambda^{®}$. Therefore, $\lambda^{®}$ is true, and, since it has been proved, it also contracts. Contradiction. While a full treatment of $\lambda^{®}$ lies beyond the scope of this paper, let me briefly indicate that its basic rot lies in the thought that, if '$\lambda^{®}$ is true and contracts' contracts, an untruth eventually follows: since the reasoning to untruth hinted at in that thought actually requires *multiple uses* of the assumption that '$\lambda^{®}$ is true and contracts' contracts, as long as *predications of contraction* (such as ''$\lambda^{®}$ is true and contracts' contracts') *can also fail to contract* untruth does not follow from that assumption. Thanks to an anonymous referee for asking about revenge.

Lisbon) and at the Veritas Pluralism, Language and Logic Workshop (Yonsei University); in 2016, at the BA Logic Group WIP Seminar (University of Buenos Aires). I'd like to thank all these audiences for very stimulating comments and discussions. Special thanks go to Eduardo Barrio, Aurélien Darbellay, John Horden, Hannes Leitgeb, Dan López de Sa, José Martínez, Julien Murzi, Francesco Paoli, Nikolaj Pedersen, Lucas Rosenblatt, Sven Rosenkranz, Ricardo Santos, Célia Teixeira, Pilar Terrés, Zach Weber, Jeremy Wyatt, David Yates and two anonymous referees. Special special thanks go to David Ripley, whose open-minded and perceptive feedback throughout the years and the continents has helped me in developing the view I present in the paper. I'm also grateful to the guest editors Riccardo Bruni and Shawn Standefer for inviting me to contribute to this special issue and for their extraordinary support and patience throughout the process, which very fittingly involved a few revisions on my part (I'm also indebted to them for this pun). Thanks guys. At different stages, this study has been funded by the Marie Skłodowska-Curie Intraeuropean Research Fellowship 301493 *A Noncontractive Theory of Naive Semantic Properties: Logical Developments and Metaphysical Foundations* and by the FCT Research Fellowship IF/01202/2013 *Tolerance and Instability: The Substructure of Cognitions, Transitions and Collections*. Additionally, the study has been funded by the Russian Academic Excellence Project *5-100*. I've also benefited from support from the Project CONSOLIDER-INGENIO 2010 CSD2009-00056 of the Spanish Ministry of Science and Innovation *Philosophy of Perspectival Thoughts and Facts*, from the Project FFI2012-35026 of the Spanish Ministry of Economy and Competition *The Makings of Truth: Nature, Extent, and Applications of Truthmaking*, from the Project FFI2015-70707-P of the Spanish Ministry of Economy, Industry and Competitiveness *Localism and Globalism in Logic and Semantics* and from the FCT Project PTDC/FER-FIL/28442/2017 *Companion to Analytic Philosophy 2*.

References

1. Beall, J.C. (2009). *Spandrels of Truth*. Oxford: Oxford University Press.
2. Belnap, N. (1982). Gupta's rule of revision theory of truth. *Journal of Philosophical Logic*, *11*, 103–116.
3. Berra, Y. (1998). *The Yogi Book*. New York: Workman Publishing Company.
4. Bolzano, B. (1837). *Wissenschaftslehre.* Vol. II. Sulzbach: Seidel.
5. Bolzano, B. (1851). *Paradoxien des Unendlichen*. Leipzig: Reclam.
6. Brady, R. (2006). *Universal Logic*. Stanford: CSLI Publications.
7. Correia, F., & Schnieder, B. (Eds.) (2012). *Metaphysical Grounding: Understanding the Structure of Reality*. Cambridge: Cambridge University Press.
8. Dedekind, R. (1872). *Stetigkeit und irrationale Zahlen*. Braunschweig: Vieweg.
9. Dedekind, R. (1888). *Was sind und was sollen die Zahlen?* Braunschweig: Vieweg.
10. Field, H. (1994). Deflationist views of meaning and content. *Mind*, *103*, 249–285.
11. Field, H. (2008). *Saving Truth from Paradox*. Oxford: Oxford University Press.
12. French, R. (2016). Structural reflexivity and the paradoxes of self-reference. *Ergo*, *3*, 113–131.
13. Gentzen, G. (1934). Untersuchungen über das logische Schließen I. *Mathematische Zeitschrift*, *39*, 176–210.
14. Girard, J.-Y. (1995). Linear logic: Its syntax and semantics. In Girard, J.-Y., Lafont, Y., Regnier, L. (Eds.) *Advances in Linear Logic* (pp. 1–42). Cambridge: Cambridge University Press.
15. Gupta, A. (1982). Truth and paradox. *Journal of Philosophical Logic*, *11*, 1–60.
16. Gupta, A., & Belnap, N. (1993). *The Revision Theory of Truth*. Cambridge MA: MIT Press.
17. Heck, R. (2012). A liar paradox. *Thought*, *1*, 36–40.
18. Hegel, G. (1813). *Wissenschaft der Logik.* Vol. I, Book 2. Nuremberg: Schrag.
19. Herzberger, H. (1970). Paradoxes of grounding in semantics. *The Journal of Philosophy*, *67*, 145–167.
20. Herzberger, H. (1982). Naive semantics and the liar paradox. *The Journal of Philosophy*, *79*, 479–497.
21. Herzberger, H. (1982). Notes on naive semantics. *Journal of Philosophical Logic*, *11*, 61–102.
22. Horwich, P. (1998). *Truth.* 2nd edn. Oxford: Clarendon Press.
23. Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy*, *72*, 690–716.
24. López de Sa, D., & Zardini, E. (2007). Truthmakers, knowledge and paradox. *Analysis*, *67*, 242–250.
25. López de Sa, D., & Zardini, E. (2011). No-no. Paradox and consistency. *Analysis*, *71*, 472–478.

26. Łukasiewicz, J. (1906). Analiza i konstrukcja pojęcia przyczyny. *Przegląd filozoficzny*, *9*, 105–179.
27. Mares, E., & Paoli, F. (2014). Logical consequence and the paradoxes. *Journal of Philosophical Logic*, *43*, 439–469.
28. McGee, V. (1991). *Truth, Vagueness, and Paradox*. Indianapolis: Hackett.
29. Paoli, F. (2002). *Substructural Logics: A Primer*. Dordrecht: Kluwer.
30. Priest, G. (2006). *In Contradiction.* 2nd edn. Oxford: Oxford University Press.
31. Reinach, A. (1911). Zur Theorie des negativen Urteils. In Pfänder, A. (Ed.) *Münchener Philosophische Abhandlungen. Theodor Lipps zu seinem sechzigsten Geburtstag gewidmet von früheren Schülern* (pp. 196–254). Leipzig: Barth.
32. Schmaltz, T. (Ed.) (2014). *Efficient Causation*. Oxford: Oxford University Press.
33. Shapiro, L. (2015). Naive structure, contraction and paradox. *Topoi*, *34*, 75–87.
34. Smith, B. (1989). Logic and the *Sachverhalt*. *The Monist*, *72*, 52–69.
35. Standefer, S. (2016). Contraction and revision. *The Australasian Journal of Logic*, *13*, 58–77.
36. von Neumann, J. (1929). Über eine Widerspruchfreiheitsfrage in der axiomatischen Mengenlehre. *Journal für die reine und angewandte Mathematik*, *160*, 227–241.
37. Weber, Z. (2014). Naive validity. *The Philosophical Quarterly*, *64*, 99–114.
38. Wittgenstein, L. (1921). Logisch-philosophische Abhandlung. *Annalen der Naturphilosophie*, *14*, 185–262.
39. Yablo, S. (1982). Grounding, dependence, and paradox. *Journal of Philosophical Logic*, *11*, 117–137.
40. Yaqūb, A. (1993). *The Liar Speaks the Truth*. Oxford: Oxford University Press.
41. Zardini, E. (2008). Truth and what is said. *Philosophical Perspectives*, *22*, 545–574.
42. Zardini, E. (2011). Truth without contra(di)ction. *The Review of Symbolic Logic*, *4*, 498–535.
43. Zardini, E. (2012). Truth preservation in context and in its place. In Dutilh-Novaes, C., & Hjortland, O. (Eds.) *Insolubles and Consequences* (pp. 249–271). London: College Publications.
44. Zardini, E. (2013). It is not the case that [*P* and 'It is not the case that *P*' is true] nor is it the case that [*P* and '*P*' is not true]. *Thought*, *1*, 309–319.
45. Zardini, E. (2013). Naive *modus ponens*. *Journal of Philosophical Logic*, *42*, 575–593.
46. Zardini, E. (2014). Naive truth and naive logical properties. *The Review of Symbolic Logic*, *7*, 351–384.
47. Zardini, E. (2014). És la veritat una mentida? Perspectives sobre les paradoxes semàntiques. *Anuari de la Societat Catalana de Filosofia*, *25*, 181–202.
48. Zardini, E. (2015). Getting one for two, or the contractors' bad deal. Towards a unified solution to the semantic paradoxes. In Achourioti, T., Fujimoto, K., Galinon, H., Martínez, J. (Eds.) *Unifying the Philosophy of Truth* (pp. 461–493). Dordrecht: Springer.
49. Zardini, E. (2015). The opacity of truth. *Topoi*, *34*, 37–54.
50. Zardini, E. (2015). ∀ and ω. In Torza, A. (Ed.) *Quantifiers, Quantifiers, and Quantifiers: Themes in Logic, Metaphysics, and Language* (pp. 489–526). Dordrecht: Springer.
51. Zardini, E. (2017). Further reflections on sentences saying of themselves strange things. *Logic and Logical Philosophy*, *26*, 563–581.
52. Zardini, E. (2018). Closed without boundaries. *Synthese*. Forthcoming.
53. Zardini, E. (2018). Generalised Tarski's thesis hits substructure. In Kellen, N., Pedersen, N., Wyatt, J. (Eds.) *Pluralisms in Truth and Logic*. Basingstoke: Palgrave MacMillan. Forthcoming.