

---

## II. LOGIC AND SCIENCE AGAINST THE BACKGROUND OF FICTION

---

### 8. Modal Fictionalism and Self-Reference

*Blagovest S. Mollov, bmollov@phls.uni-sofia.bg*  
(Professor, Dr. Sc., Department of European Studies,  
University of Sofia St. Kliment Ohridski)

**Abstract:** Modal fictionalism – as a deflationary meta-ontological proposal of how to retain the semantic benefits of Lewisian modal realism without its ontological costs – has usually been attacked for either failing to provide modal realist benefits in full or/and for incurring its own extravagant epistemological and ontological costs. However, the most fatal flaws the view manifests are of a self-referential character, rendering it circular, paradoxical, or inconsistent. The paper discusses various self-reference objections and argues that none of the advanced modal fictionalist response strategies – meta-fictional, primitivist, paraphrastic, indexical, etc. – can succeed, as any fictionalization is bound to collapse into incoherence, due to inherent, unfixable structural defects shared by all types of fictionalism.

**Keywords:** modal fictionalism, self-referentiality, primitivism

Self-reference (symmetry) arguments against modal fictionalism (MdF) seem to be the closest analytic metaphysics has got to that unicorn species of “knock-down” arguments, the elusiveness of which has always been the plight – and pride – of philosophy. I.e., if sound, the self-reference arguments would show MdF to be incoherent, thereby refuting it conclusively and irrevocably. In what follows, I sketch the reasons why they are, and would.

Like any fictionalism, modal fictionalism proposes a way to enjoy the benefits of a certain theory – in this case, the technically advantageous possible-worlds framework – without paying the asking ontological price that its adoption seemingly requires, viz., ontological commitment to possible worlds as either Plantingan abstract or Lewisian concrete entities by treating claims *about possible worlds* as fictional (i.e.,

literally false). And, as with any other fictionalism, the truth-conditions of the given class of statements, here modal statements, get relativized to the respective fiction, “the fiction of possible worlds” (PW) (the fiction of, for example, Lewisian modal realism).

Thus, modal claims are supposed to be understood as having an implicit premise “In fiction PW, p”, as claims about the contents of the PW fiction, so that a modal claim p would be true if and only if it holds *according to the fiction PW*. I.e., “Possibly, p” standardly rendered as “There is a possible world where p” is taken to mean “According to the fiction of possible worlds (PW), there exists a world where p”. Or,  $\Diamond p$  is to be analyzed as  $F_{PW} (\exists w (at\ w, p))$ , respectively,  $\Box p$  as  $F_{PW} (\forall w (p@w))$ , where “ $F_{PW}$ ” is the fiction operator (“according to the fiction PW”), which appears to be plausible enough and to have the added theoretical virtue of being in tune with conventional intuitions about the truth-conditions and ontological commitments of fictional statements.

Schematically, any fictionalism about a given domain D (modality, morality, mathematical or scientific entities) incorporates three core components: (i) a metaphysical (or a meta) thesis that propositions of D-discourse are, literally construed, false (or truth-valueless) because the entities (e.g., possible worlds, moral properties, abstract objects, etc.) they refer to do not exist; (ii) an instrumentalist claim that D-framework is nevertheless theoretically useful and, so, should and could be used by adopting (iii) a suitable paraphrase schema: “SD (a D-discourse statement) is true, if and only if, according to the fiction F, SD\* (SD\* being the paraphrase of SD in the fiction F)”. Modal fictionalism demonstrates the same structure: it rejects the existence of possible worlds, to which possible-worlds discourse ontologically commits, and proposes a translation schema:  $\Diamond p \Leftrightarrow PW \vdash \exists w (p@w)$  ( $\Box p \Leftrightarrow PW \vdash \forall w (p@w)$ ) (where PW is the meta-fictional operator, “according to the fiction of possible worlds”)) that would allow the beneficial use of possible-worlds framework without its unwanted ontological commitments.

But while all fictionalisms face reflexive pressure, modal fictionalism is affected by more and deeper issues than its peers because modality is, what can be called, “globally structurally constitutive”, i.e., modal concepts are built into, presupposed by, the explanatory, inferential and evaluative practices of any theorizing whatsoever, including theorizing about modality. Unlike the localized, circumscribable, isolable domains with proprietary entities of, say, mathematics (numbers, sets, functions,

etc.) or morality (moral properties, facts, values, etc.), modal notions, being conditions for the possibility of theorizing in general, are inextricable, so that modal fictionalism – as a meta-*theory* about modality – has no external non-modal perspective on the modal domain.

Since the whole methodological strategy of fictionalism depends on the availability of a neutral meta-language, an external standpoint from which to treat the target discourse as fictional, i.e., to reject the ontological commitments of the discourse and explain the pragmatic legitimacy of preserving its use, the lack of such a standpoint in the case of modality means that any attempt to fictionalize the latter will generate ruinous self-reference problems, as the target discourse would inevitably be reintroduced at the meta-level. These do not arise for mathematical, moral or other types of fictionalism, for none of their target discourses are similarly globally structurally constitutive; hence, those fictionalisms can enjoy the required external neutral perspective for the non-circular fictionalization of their respective target discourses and be formulated, without incoherence, using conceptual resources that do not belong to the target discourses (generally by using a physicalist meta-language).

Thus, modal fictionalism seems to suffer from a structural, hence insurmountable, fatal flaw. For, if modal concepts are constitutive prerequisites for the formulation, articulation, inferential mechanics, and assessment of any theory, including theories about modality, any meta-language powerful enough to serve as an adequate analysis of modality, as modal fictionalism claims to provide, will necessarily instantiate the conceptual resources of the modal domain, thereby failing the “neutrality requirement” and collapsing into self-referential incoherence.

Given that the constitutive character of the modal discourse cannot be challenged, modal fictionalism’s range of response options is severely limited: it would have to either concede circularity<sup>1</sup>, giving up its reductive ambitions, or commit to primitivism about the meta-linguistic “according to PW” operator or the fictionalist paraphrase schema, giving

---

<sup>1</sup> As Rosen explicitly did: “the background ontology and ideology in which the fiction is described is already deeply modal.” in Gideon Rosen. “A Problem for Fictionalism About Possible Worlds”. *Analysis* 53(2) (1993): 75. Cf. also Rosen: “The principal worry is that the fictionalist’s paraphrase may surreptitiously presuppose the very modal notions it is supposed to explain.” in Gideon Rosen. “Modal Fictionalism”. *Mind* 99(395) (1990): 334.

up (some of) its explanatory ambitions<sup>2</sup>.

Whatever objections one may have against primitivism, it does appear to be the better, or the only, option: by treating the “according to PW” operator, usually understood as a kind of entailment (a modal notion: necessary consequence), as a non-modal semantic primitive (a brute relation between a fiction (a set of propositions) and a proposition), circularity would be avoided. The strategy is expected to block self-reference objections as well as concerns about the related metalanguage infinite regress that assuming the analyzability of the operator may trigger. The fictionalist primitive “according to” relation between a fiction and a proposition also parallels the precedent strategy of “ersatz” modal realism, similarly positing primitive operators or primitive representation relations between abstract entities and possibilities, which is supposed to further legitimize the move<sup>3</sup>.

The question here is whether such a primitive semantic operator can be fully detached from modal involvement. Rosen suggests that it need not be<sup>4</sup>, at least not fully, since apparently some modal notions cannot be reductively analyzed and all fictionalism needs is a minimal, “thin” modal notion, similar to “consequence”, with which to define “truth-in-a-fiction”. So, “according to the fiction PW, p” would get construed as “p is a consequence of the fiction PW”, where “consequence” is taken as a primitive, non-reductive concept. The concept of “consequence” need not be completely non-modal, but, as Rosen puts it, non-“objectionably modal”<sup>5</sup>, meaning non-committal to the ontology of possible worlds, i.e., some modal concepts are admissible as long as they

---

<sup>2</sup> Cf. Rosen: “One possible response to this dilemma is to insist that ‘according to PW’ is a theoretical primitive for the modal fictionalist. It is not to be explained in more fundamental terms.” in Gideon Rosen. “A Problem for Fictionalism About Possible Worlds”, 76; cf. also Gideon Rosen. “Modal Fictionalism Fixed”. *Analysis* 55(2) (1995): 70.

<sup>3</sup> Cf. Rosen: “[if the ersatzer] is entitled to help himself to this ideology without further explanation, then the fictionalist is entitled to help himself to ‘according to PW’ on the same terms” Gideon Rosen. “A Problem for Fictionalism About Possible Worlds”, 76.

<sup>4</sup> Cf. Rosen: “The fictionalist need not deny that there are primitive modal facts. What matters is only that these facts not involve commitment to concrete possible worlds.” in Gideon Rosen. “Modal Fictionalism”, 338.

<sup>5</sup> Cf. Rosen: “There is nothing objectionably modal in the claim that one sentence follows from others in virtue of their content.”, *ibid.*, 339.

are “ontologically innocent”, bear no unwanted ontological commitments. Thus, the primitivist strategy supposedly blocks circularity not by eliminating all modality from fictionalist metalanguage, but by denying that the circularity in question is vicious as long as the modal notions at the primitive level are not ontologically committal.

That, however, is highly questionable as the claim that “to be a consequence of” or “follow from” can be ontologically non-committal and, hence, can be used in the definition of the “according to PW” operator, goes against the standard – substantively, rather than “thinly”, modal – possible-world semantics of “consequence”: “According to PW,  $p$ ” = “ $p$  is a consequence of PW” is exactly the “objectionable” “in every possible world where PW is true,  $p$  is true”. For what else can “to be a consequence of” mean? There simply seems to be no way to de-ontologize modal notions, no matter how primitive they are construed to be. So, the primitivist circularity will, after all, be vicious and ineluctably so.

While primitivists are fully aware<sup>6</sup> that if this possible-worlds meaning were definitionally constitutive of “consequence” or were, in some sense, metaphysically prior, the viciousness of the proposed circularity follows indeed; they deny that this is the case. I.e., they reject what they take to be the unfounded universal assumption that “a proposition  $p$  follows from (is a consequence of) a fiction  $F$ ” means and could only mean “ $p$  is true in every possible world where  $F$  is true”. They insist that, actually, “consequence” should be understood as a primary primitive inferential relation among “contents”, propositions, sentences, with the possible-worlds interpretation being just one, but not the only, representation (model) of it, commonly the one privileged by modal realists. That original “consequence” relation bears no implicit or explicit ontological commitments, precisely what it being merely inferential means<sup>7</sup>: “ $p$  follows from PW” expresses only a claim about the inferential structure of a set of propositions (PW, the fiction of possible worlds) rather than refer to or quantify over realistically interpreted objects of a certain kind, “possible worlds”. Analogously to the standard logical notion of “semantic consequence” whose use does not commit to any particular domain of objects (or even to any objects whatsoever) over

<sup>6</sup> Cf. Rosen: “The notion of consequence is often explained in modal terms, and if that were unavoidable the fictionalist proposal would be circular.”, *ibid.*, 340.

<sup>7</sup> Cf. Rosen: “Inference is not an ontological relation. It is a relation among propositions, not among entities in the world.”, *ibid.*, 341.

which models range, primitivist “modal consequence” is said to be usable in a similar ontologically neutral way: it can be safely represented in terms of possible worlds without being semantically constituted by such, modal-realistically interpreted, entities.

That can be a controversial response, though, since even if a primitive modal operator of that kind, like its supposed logical analogue, does not commit to a particular domain of objects, it does commit to a space of possible domains, of possibilities, what those “ontologically neutral” arbitrary set-theoretic models actually are, which rather than being “ontologically neutral” seem to commit to, at least, a “robust” ontology of abstract objects and to an irreducible modal background in which those domains exist as possibilities that can strongly suggest a realist modal ontology. Since apparently the most productive way to refer to that modal space of possibilities and its locations is “possible worlds”, the very reason modal fictionalism wants to keep the possible-worlds framework in use, it is not clear how the primitivist move expects to avoid circularity. The primitivist proposal of treating “possible worlds” purely instrumentally as a representational device, the way geometrical coordinates or maps are supposedly used without being ontologically committing, requires additional argumentation, as it depends on metaphysical (nominalist, instrumentalist, fictionalist) assumptions that are highly contentious, and a demonstration of how the strategy can be successfully applied to actual cases in first-order modal metaphysics. For it would be very hard otherwise to convince the Quineanly-inclined orthodoxy that one can use a theory that quantifies over a certain kind of objects, in this case, abstract objects (which “possible structures”, models, sets, modal space coordinates/regions/sections/points or other supposed “mere semantics tools” unquestionably are); treat its propositions, and the theory itself, as truth-evaluable, and somehow still deny that one is committed to the existence of those objects<sup>8</sup>. So, it appears, for now, viciousness has not been averted.

---

<sup>8</sup> As Lewis, with characteristic bluntness, remarks, targeting the Tarskian idea of a “metaphysically innocent” “consequence”: “You can call it ‘consequence’ if you like, but it walks and quacks exactly like necessity.” in Lewis, David K., *On the Plurality of Worlds* (Oxford: Blackwell, p. 153). Rosen is aware, but, having different metaphysical assumptions, disagrees: “It may be complained that consequence is simply necessity by another name.” in Gideon Rosen. “Modal Fictionalism”, 340.

Primitivists may still insist, however, that a properly drawn and understood distinction between levels of modality, namely, between object-level modality (necessity, possibility as applied to object-language propositions) and meta-level modal concepts (such as “consequence”, “entailment”, “truth-in-a-fiction” used in theories about object-level modality) can easily settle the issue.

Since modal fictionalism employs meta-level rather than object-level modal concepts, as all theories, including modal realism, do by necessity, the apparent circularity in its analysis of modality is claimed to be benign or “merely structural”, “explanatory”. The obvious issue with the proposed distinction is that there does not seem to be any ground for distinguishing the modal sense or force of a “meta-level” consequence, entailment, etc., from object-level modal notions. Necessity does not seem to vary with levels. Even if a ground for the distinction existed, it would have to be a metaphysical, not terminological one. But then, if both types of modality are grounded in modal properties, facts, etc. of reality, primitivist modal fictionalism has failed to avoid substantive modal ontological commitments, just re-labelled them. And finding a principled metaphysical ground for the meta-/object-level distinction of the required sort - a metaphysical distinction between “acceptable” (“thin”, “lightweight”) and “objectionable” (“robust”, “heavyweight”) modal facts - is even more hopeless a task.

Furthermore, once primitivist modal fictionalists admit primitive modal facts (“thin” and “robust”), the need for fictionalism disappears, as the posited possible-world fiction becomes redundant, explanatorily idle: “Necessarily, p” will be true because of the relevant modal primitive facts, not because of the PW fiction (“Necessarily, p” because p is a primitive consequence of PW). Modal fictionalism will have lost its distinctive theoretical identity and value, collapsing into modal primitivism. The difference between the two is merely notational and not worth sustaining, as primitivist modal fictionalism offers no additional explanatory resources or theoretical virtues.

This questionability of the motivation for modal fictionalism arises even if, miraculously, the primitive semantic operator “according to PW” can be made independent of “objectionable” modal notions and ontological commitments, for modal fictionalism would have to explain why we ought to bother replacing our standard and more intuitive commitment to (primitive) modal properties and facts with one to primitive

semantic facts. Rosen acknowledges that the “according to PW...” operator is unintuitive and a very unlikely candidate to be considered primitive (unanalyzable), though arguably, the naturalness or attractiveness of primitives is a variable matter of metaphysical or methodological tastes that defies uniform ranking.<sup>9</sup>

Actually, the motivation for modal fictionalism can dramatically drop at the very point of opting for a primitive “according to PW” operator, thereby abandoning the more ambitious explanatory aim of providing a fully reductive account of modal discourse in non-modal (fictional) terms. Without its reductive aim, modal fictionalism loses every theoretical advantage it initially appeared to have over modal realism or modal primitivism. Again, if one is ready to concede that modality is primitive anyway, why bother with the fiction in the first place? If ready to commit to primitive irreducible modal notions, why not take “possible worlds” as primitive too? There do not seem to be principled grounds on which to decide which modal notions to keep in the primitive base and which not. Or better yet, unburden what essentially is a modal primitivist theory of the extra machinery of the fictional operator, which is theoretically idle anyway, since non-reductive modal fictionalism merely adds a flimsy fictionalist coating over primitive modal facts that are actually doing all the real explanatory work, instead of grounding modal facts in the fiction as its early adopters had hoped. It does appear, however, that the less theoretically ambitious modal fictionalism is – retreating to being solely a specific severely deflationary semantic theory of modal discourse – the higher its chances of a theoretical future<sup>10</sup>.

Though one may sceptically ask, “What future?” For if, in trying to avoid the collapse into modal primitivism, modal fictionalism adopts some kind of safe conventionalism about its principle, it will have *no* theoretical future, as that final shift would completely gut it of any hope

---

<sup>9</sup> Rosen *tu quoque* fallaciously notes that the realist competitors to fictionalism, implicitly referring to Lewis’s analysis, have similarly failed to provide a satisfactory account of “according to a fiction...” operator, which, even if true, is only a polemically understandable move with no substantive justificatory value (for only in very specific urgent practical contexts can the failure of alternatives be considered a viable argument for a view or a course of action), *ibid.*, 345-6.

<sup>10</sup> Rosen explicitly concedes, perhaps prolonging the theoretical lifespan of fictionalism but depriving that prolongation of reasons to matter: “The fictionalist’s proposal should not be understood as a deep metaphysical explanation of modality.”, *ibid.* 341.

of or claim to capturing substantive objective truth about modality, reducing it to a pragmatic linguistic choice of a particular expressive, convenient or figurative, representational idiom. While the conventionalist shift appears to be the only viable way for modal fictionalism to avoid incoherence, it would amount to a theoretical suicide.

To end on a more “positive” therapeutic note, that existential predicament is not exclusive to *modal* fictionalism, due solely to its character of being “modal”, i.e., due to the specificity of its subject-matter, the mentioned global constitutivity of modality; it affects all types of fictionalism – moral, mathematical, scientific, etc. – insofar as they are *fictionalisms*. Fictionalism, as a general strategy, is inherently an irredeemably and irredeemably incoherent theoretical stance.

Modal fictionalism’s conventionalist shift was forced by the self-reference pressure to specify the modal status of modal fictionalism’s biconditional translation schema: (SMdF)  $\Diamond p \Leftrightarrow PW \vdash \exists w (p@w)$  (According to the fiction of possible worlds  $PW$ ,  $\exists w (p \text{ at } w)$ ). The dilemma (Hale’s dilemma<sup>11</sup>) poses a straightforward question: Is (SMdF) necessarily true or contingently true? The first horn of the dilemma, (SMdF) is necessarily true, results in circularity (using the fictionalist schema to justify the fictionalist schema) or collapse into modal primitivism (to avoid vicious circularity, declaring (SMdF) to be a primitive necessary truth, i.e., conceding an “unobjectionable” modal primitive and thereby presupposing independent non-fictionalist sources of modal truths). If, dilemma’s second horn, (SMdF) is contingently true, apart from the blow to the schema’s authority (it could have failed), circularity follows again, as we will need to rely on independent modal facts and notions to explain the contingency of (SMdF), i.e., why it holds in the actual world but might not have. The conventionalist move, viz., further deflating (SMdF) to be a contingent, conventionally adopted semantic rule, is the direct, theoretically suicidal response to the dilemma.

Now, an analogous dilemma applies to all fictionalisms in general. If for any discourse  $D$ , fictionalism about that discourse claims:

(SFD)  $p \Leftrightarrow F \vdash p^*$  “ $p$  (a  $D$ -discourse proposition) is true, if and only if, according to the fiction  $F$ ,  $p^*$  ( $p^*$  being the paraphrase of  $p$  in  $F$ ),

where that involves an additional meta-claim that  $D$ -discourse state-

<sup>11</sup> Hale, Bob. 1995. “Modal Fictionalism: A Simple Dilemma”, *Analysis* 55(2): 63–67.

ments are truth-valueless or false because they have no real referents, then, the dilemma question would be:

Is (SFD) fictionally true or non-fictionally true?

If Horn I, (SFD) is fictionally true, i.e., the truth of the schema (SFD) itself is explained as truth-in-the-fiction F (as self-grounded), then circularity follows (justifying the fictionalist schema by the fictionalist schema) and, more importantly, (SFD) gets relativized to a fiction, which would eliminate its apparently intended normative authority as an alternative to realism about D. If (SFD)'s truth is internalized to a fiction, be it a meta-fiction (about ontological language), it can no longer claim to be such an alternative since it would carry no more normative "truth force" than "Holmes lives on 221B Baker Street" or "Emma Bovary revelled in horse riding", i.e., fictionalism would amount to no more than pretending that realism about D is false, an arbitrary meta-language game of make-believe.

The attempt to make Horn I acceptable by supporting it with the contention that all (ontological) theories are to be treated as fictions, so that every theoretical truth would be truth-in-a-fiction, simply relocates the issue, for, now, would that contention itself be fictionally or non-fictionally true? Moreover, since "fiction" is uniquely defined by truth-in-fiction vs. truth-in-reality distinction, dependence on a practice of non-fictional language/literary community that distinguishes fiction from non-fiction; imaginative pretense (on the part of the non-fictional audience), and non-truth-assertiveness (on the part of the non-fictional), that attempt of legitimizing Horn I would require one to be radically revisionary about the semantic and conceptual content of "fiction", collapsing all the distinctive definitional distinctions of "fiction", thereby literally leaving the bounds of intelligibility, and eliminating the whole point – and intelligibility – of fictionalism.

If Horn II, (SFD) is non-fictionally true, then the fictionalist thesis would no longer be uniformly valid, allowing for the non-fictional truth of at least some theoretical meta-claims. But once that exception has been made, if literal, non-fictional truthful theses about D are admitted as possible, fictionalism would have to explain why it denies the same literal truthfulness to statements within the D-discourse and alternative realist meta-claims about D. Given the additional unintuitive theoretical burden of the fictionalist explanatory apparatus, the initial motivation

(epistemic skepticism, ontological parsimony, radical anti-realism etc.) that global, fully reductive fictionalism may have attracted disappears.

The dilemma seems to be devastating: if fictionalism is merely fictionally true, it would provide no reason to alter our alternative realist conceptual and linguistic practices; if non-fictionally true, it would no longer be fictionalism, the deflationary radical view it could have at least tried to be, but a modified realism with a contestable fictionalist claim about a restricted discourse.

The only benefit fictionalism may have is drawing our attention to the potential limits and contingency of our ontological realist languages and frameworks, pointing to the possibility of alien ones, but since no specific aspects and conditions of that (possibility for) alienness are explored, and neither are the ones of a meta-language to identify framework limitations and distinguish alien and non-alien frameworks, that particular benefit of engaging fictionalism is only of a gestural, rather vaguely implicit nature and might be eventually completely lost in the ongoing shift from radical to conventionalist fictionalism.