Essences, Heuristics, and Metaphysical Illusions

Timothy Williamson
(penultimate version of paper to appear in *Metaphysics*)

Abstract: The paper is a critique of Kit Fine's central argument in 'Essence and Modality' for the hyperintensionality of essence-attributing operators, from the premise that it is essential to Socrates that he is Socrates but not that he belongs to {Socrates}. Similar arguments can be given about what is essential to natural numbers, but are provably unsound. The proof of unsoundness exploits the category of singular terms with a compositionally complex semantics that are nevertheless directly referential, such as '7 + 1'; they must not be confused with definite descriptions. The proof depends on the standard logic of identity, not on any intensionalist assumption. The errors in our pre-theoretic essentialists judgments are explained by our reliance on an efficient but fallible heuristic. The analogous premises of Fine's argument are generated by the same heuristic and so are untrustworthy. More generally, it is suggested, hyperintensionalist theorizing is guilty of overfitting dodgy data. The paper also notes a strand of Fine's original article that restricts its claims to the intelligibility rather than truth of hyperintensional essentialism; its intelligibility is not contested.

1. *Retrospective*

I first encountered a version of Kit Fine's seminal paper 'Essence and Modality' in 1992, when he presented it to an informal discussion group hosted by David Charles in his rooms at Oriel College, Oxford. It made a powerful impact. We came away with the impression that Fine was offering potentially devastating counterexamples to a widely held, Kripke-inspired modal approach to understanding essence, cases where one proposition but not another is essential to an object, even though the two propositions are necessarily equivalent. In particular, it is essential to Socrates that he is Socrates but inessential to him that he belongs to singleton Socrates, the set {Socrates}, even though, necessarily, he is Socrates if and only if he belongs to singleton Socrates. Thus, since essence cuts finer than modality, the first is irreducible to the second. That is how the paper has generally been received since it was published (Fine 1994). As a result, it has inspired the subsequent more general search for finer-grained, hyperintensional, metaphysical structure, which has clearly been one of the most significant trends in metaphysics over the past thirty years. There is even talk of a 'hyperintensional revolution' to match the 'intensional revolution' of the 1960s (Nolan 2014).

2. *Conceptual and metaphysical interpretations*

Rereading 'Essence and Modality' in 2024, I was surprised to find that its official claims are much more cautious than I had remembered, and than the received interpretation implies. For instance, we read:

> Nor is it critical to the example [about Socrates and singleton Socrates] that the reader actually endorse the particular modal and essentialist claims to which I have made appeal. All that is necessary is that he should recognize the intelligibility of a position which makes such claims. For any reasonable account of essence should not be biased towards one metaphysical view rather than the other. It should not settle, as a matter of definition, any issue which we are inclined to regard as a matter of substance. (5)[1]

On this conciliatory interpretation, the point of the example is not to show that modal accounts of essence are *false*, but merely to show that they are *synthetic* rather than analytic, in some relevant sense.[2]

A similar point may be intended in this passage:

> For it seems to be possible to agree on all of the modal facts and yet disagree on the essentialist facts. But if any modal criterion of essence were correct, such a situation would be impossible. (8)

Of course, if two people agree on all modal facts but not on some essentialist facts, it follows trivially that some essentialist facts are not modal facts. But that argument is dialectically ineffective unless one has already ruled out the view that essentialist facts *are* modal facts 'in disguise'. But the passage avoids such question-begging when read as suggesting that

philosophers can agree in all *explicitly* modal discourse while disagreeing in some *explicitly* essentialist discourse. One might conclude that explicitly essentializing terms are not *synonymous* with explicitly modal alleged analyses.

Here is a third passage:

Given the insensitivity of the concept of necessity to variations in source [whose essence yields necessity], it is hardly surprising that it is incapable of capturing a concept which is sensitive to such variation. (9)

Such a difference between the two concepts does not by itself force a difference between what they are concepts of. A theorist could in principle concede that the concept 'essential' involves a source parameter, which the concept 'necessary' lacks, while still arguing that the parameter is a redundancy in thought that makes no difference in reality.[3] Not all differences between concepts project onto differences between what they are concepts of.

On the envisaged concessive view that 'Essence and Modality' in this mood permits (but does not endorse), essentiality is in fact necessity, even though the *concept* 'essential' is not the *concept* 'necessary'; the genuine cognitive differences in our uses of the two words need not reflect differences in what they refer to. Likewise, gold is in fact the element with atomic number 79, even though the *concept* 'gold' is not the *concept* 'the element with atomic number 79'; the genuine cognitive differences in our uses of the word and phrase need not reflect differences in what they refer to. Again, on the pretence that the good is in fact what maximises utility, the *concept* 'good' is still not the *concept* 'what maximises utility'; the genuine cognitive differences in our uses of the word and the phrase do not prove a difference in what they refer to, whatever G.E. Moore's 'open question argument' may be intended to show.

A similar caution appears in the paper's treatment of a subsidiary theme: the analogy between essence and necessity, on one side, and definition and analyticity, on the other. Just as an essence is overtly the essence of *something*, so too a definition is overtly the definition of *something*. By contrast, neither necessity nor analyticity is overtly so sourced in anything, though there is the metaphysical theory that all necessity is sourced in essences of things, and the analogous theory that all analyticity is sourced in definitions of things; Fine calls the denial of the latter theory 'holism'. Here too, he is surprisingly concessive:

However, just as in the essentialist case, the important issue concerns intelligibility rather than truth. We want to know if there could be a genuine difference of opinion as to whether 'man' is correctly definable as 'bachelor or husband' or as to whether some form of holism is correct; and when the point is put in this way, it seems hard to see how it could be denied. (11)

He also echoes the discussion of agreement and disagreement:

For just as it appeared to be possible to agree on the modal facts and yet disagree on the essentialist facts, so it appears to be possible to agree on the facts of analyticity and yet disagree on the facts of meaning. (11)

As before, some interpretation is required. If two people agree on all facts about analyticity but not on some facts about definitions, it follows trivially that some facts about definitions are not facts about analyticity. But that argument is dialectically ineffective unless one has already ruled out the view that facts about definitions *are* facts about analyticity 'in disguise'. The passage avoids such question-begging when read as suggesting that philosophers can agree in all discourse *explicitly* about analyticity while disagreeing in some discourse *explicitly* about definitions. One might conclude that some terms of the latter discourse are not *synonymous* with alleged analyses in terms of the former discourse.

In such passages, 'Essence and Modality' presents itself as unambitiously defending just the *intelligibility* of the striking metaphysical claims for whose *truth* the paper is standardly read as providing robust arguments. That the paper achieves the modest goal of vindicating the intelligibility of those metaphysical claims is hardly in doubt. Obviously, by ordinary standards, many competent, reasonable, unconfused speakers of English—for instance, Kit Fine—can understand and even assent to statements like 'Socrates is necessarily but not essentially a member of the set whose only member is Socrates'. The question is whether the metaphysical claims are true, not whether they are so much as intelligible.

Confusingly, in other passages, the paper is much less inhibited, and does address the question of truth:

> But, intuitively, this [that Socrates essentially belongs to singleton Socrates] is not so. It is no part of the essence of Socrates to belong to the singleton. (4-5)

This is an outright claim about what is in the essence of Socrates, not just a claim about what is in our concept 'essence of Socrates'. It corresponds to the claim that hydrogen is no part of gold, not to the mere claim that our concept 'hydrogen' is no part of our concept 'gold'. The paper contains many more passages like that:

> There is nothing in the nature of a person, if I may put it this way, which demands that he belongs to this or that set or which demands, given that the person exists, that there even be any sets. (5)

> But it is not essential to Socrates that he be distinct from the [Eiffel] Tower, for there is nothing in his nature which connects him in any special way to it. (5)

> But it is no part of Socrates' essence that there be infinitely many prime numbers or that the abstract world of numbers, sets, or what have you, be just as it is. (5)

In these passages, Fine directly tells the reader what is in the essence of Socrates, though he does not explain how he knows. Thus, the usual metaphysical reading of 'Essence and Modality' also has clear textual support.

This paper concerns the metaphysical view for which 'Essence and Modality' is famous, irrespective of Fine's intentions in writing it. Given uncontested background assumptions, do the proposed counterexamples indeed refute modal characterizations of essence? I will argue that although the key judgments about the examples may feel compellingly natural, they issue from a way of thinking whose outputs in similar cases are

provably false; thus, the key judgments are unsafe and not to be relied on. More generally, all the extra theoretical complications incurred by the hyperintensionality of an anti-modalist theory of essence are likely to be artefacts of *overfitting* error-infected data. In other words, it results from a methodological pathology well-known in the natural and social sciences, whereby theories are made more and more complicated to achieve an exact fit with data, undermining theorists' ability to spot outlying, potentially erroneous data points.[4]

3. *Sets and natural numbers*

Consider these three statements:

(1) It is essential to Socrates that he is Socrates.

(2) It is essential to Socrates that he belongs to {Socrates}.

(3) It is essential to {Socrates} that Socrates belongs to it.

On a Finean view, (1) is true and (2) false. For the set {Socrates} is in some sense *extraneous* to the man Socrates, while Socrates is trivially not extraneous to himself. In the same sense, Socrates is *not* extraneous to {Socrates}, for sets as normally understood are somehow constituted from their members (if any); in particular, {Socrates} is constituted from its only member, Socrates. Thus, (3) is presumably true, on the Finean view. The difference in truth-value between (2) and (3) reflects an asymmetry in constitution between sets and their members.

The Finean denial of (2) is not neutral on the metaphysics of sets. It coheres with a broadly iterative conception of sets, on which they are built up from their members stage by stage. In particular, {Socrates} is built up from Socrates, while Socrates is not built up from {Socrates}, so Socrates and {Socrates} are distinct. More generally, by the foundation (or regularity) axiom of standard Zermelo-Fraenkel set theory, every non-empty set is disjoint from at least one of its members; so, always, $x \neq \{x\}$, otherwise $\{x\}$ would not be disjoint from x. However, there are also well-developed alternative theories of non-wellfounded sets, which drop the foundation axiom and allow cases where $x = \{x\}$; such an x is called a *Quine atom*. If Socrates is a Quine atom, then he *is* the set {Socrates} and belongs to it just by belonging to himself.[5] In that case, the Finean objection to (2) might well lapse. This paper does not press such concerns. A background iterative conception of sets for the denial of (2) is plausible and widely held, independently of issues about hyperintensionality; it can reasonably be conceded, at least for the sake of argument.

Natural numbers are often, and very naturally, understood as metaphorically constructed in a similarly iterative way: a natural number is the result of starting from 0 and adding 1 as many times as needed. Such an iterative conception of natural numbers, although not uncontested, is plausible and widely held, independently of issues about hyperintensionality; it too can reasonably be conceded, at least for the sake of argument.

Given the iterative conception of natural numbers, consider these three statements:

(4) It is essential to 8 that it is 8.

(5) It is essential to 8 that it is 9 – 1.

(6) It is essential to 8 that it is 7 + 1.

On a Fine-inspired view, (4) is true and (5) false. For, in the relevant sense, the natural number 9 is *extraneous* to the natural number 8, while 8 is trivially not extraneous to itself. In the same sense, 7 is *not* extraneous to 8. On the iterative conception, 8 is somehow built up from 7, and 9 from 8; 8 is not built up from 9. Presumably, (6) is true, on this Fine-inspired view. The difference in truth-value between (5) and (6) reflects an asymmetry in constitution between natural numbers and their predecessors.

Since 1 is itself one of the natural numbers to be constructed, we can enhance fidelity to the underlying vision by substituting the successor operation s for adding 1 and the predecessor operation p for subtracting 1, so that (5) and (6) become (5*) and (6*) respectively:

(5*) It is essential to 8 that it is $p(9)$.

(6*) It is essential to 8 that it is $s(7)$.

Given the iterative conception of natural numbers, one may naturally judge that (5*) is false, just like (2) and (5), while (6*) is true, just like (3) and (6).

One might worry that p is undefined on 0, since we are concerned with operations on natural numbers, not on positive and negative integers. Similarly, one might worry that the result of subtracting a larger natural number from a smaller one is undefined. One can easily solve these difficulties by stipulating that $p(0) = 0$ and that $m – n = 0$ for $m < n$. The artificiality of such stipulations only strengthens the impression that (5) and (5*) are false.

One can strengthen the analogy between the iterative conception of sets and the iterative conception of natural numbers, and subsume the latter under the former as a special case, by adopting one of the standard ways of identifying natural numbers with pure sets. On Zermelo's way, 0 is the empty set $\{\}$ and, for any natural number n, $s(n)$ is $\{n\}$. On von Neumann's way, 0 is again the empty set and for any natural number n, $s(n)$ is $n \cup \{n\}$. Both ways preserve the desired order of constitution, which is part of their appeal. However, no such identification of natural numbers with sets is assumed in what follows.

So far, apparently, so good. But something is wrong. For 9 – 1 and 7 + 1, and $p(9)$ and $s(7)$, are all just natural numbers, indeed, they are all the same natural number, 8. All five of these numerical terms coincide in reference, so substituting them for each other in (4), (5), (6), (5*), and (6*) should make no difference to their truth-value, contrary to the Fine-inspired judgments that (4), (6), and (6*) are true while (5) and (5*) are false. What is going on?

An initial reaction might be that the semantically complex expressions '9 – 1', '7 + 1', '$p(9)$' and '$s(7)$' are not names but definite descriptions. After all, they are naturally

paraphrased in ordinary English by definite descriptions such as ‘the result of subtracting one from nine’, ‘the sum of seven and one’, ‘the predecessor of nine’, and ‘the successor of seven’. Such analyses of functional expressions in terms of predicates and definite descriptions were commonplace in logicist formalizations of mathematics. If one treats definite descriptions as quantifier phrases, one might then suspect that inter-substituting co-denoting definite descriptions within the scope of operators such as ‘it is essential to 8 that’ can fail to preserve the truth-value of sentences in which they occur, and so induce fallacies analogous to those in early critiques of quantified modal logic.

On further reflection, however, such reliance on definite descriptions represents a failure to take seriously a distinctive feature of standard mathematical notation. Mathematicians often achieve an elegantly streamlined notation by treating functional expressions as basic, for example, in both algebra and analysis. A notable case in point is *primitive recursive arithmetic*, whose theoretical significance depends on its elementary character. That is achieved in part by formulating it in a quantifier-free language, where no definite description operator is available, and formulas are read as implicit generalizations over arbitrary natural numbers. Psychologically, the use of function symbols such as ‘+’ in mathematics feels entirely natural; what feels *unnatural* and clunky is to use predicates and definite descriptions instead.

There is no semantic obstacle to treating functional expressions as primitive. In particular, we can give a quasi-homophonic recursive semantics for a functional expression in a metalanguage that extends the mathematical object-language, using the same functional expression to state the relevant semantic clause, just as we might expect for a primitive expression. Informally, for example, the reference of ‘7 plus 5’ is the reference of ‘7’ plus the reference of ‘5’, which is 7 plus 5, which is 12.

More formally and generally, we consider a formal language with both constant and variable atomic singular terms, and complex singular terms built up with function symbols, For any singular term t and assignment $\underline{a}$ of values to all variables, $\mathrm{ref}_{\underline{a}}(t)$ is the referent of t under $\underline{a}$. The case of ‘+’ is exemplary. Syntactically, whenever t_1 and t_2 are singular terms, so is ‘(’^t_1^‘+’^t_2^‘)’, where ^ is concatenation of syntactic strings. Semantically, we assume for simplicity that the domain is just the set of natural numbers. Then the natural semantic clause for ‘+’ says, for any singular terms t_1 and t_2:

$\mathrm{ref}_{\underline{a}}$(‘(’^$t_1$^‘+’^$t_2$^‘)’) = $\mathrm{ref}_{\underline{a}}(t_1) + \mathrm{ref}_{\underline{a}}(t_2)$

Analogous syntactic and semantic treatments apply to the function symbols for subtraction, the successor and predecessor operations (totally defined as above), multiplication, exponentiation, and so on. Together, these clauses recursively determine the referent (under an assignment) of arbitrarily complex singular terms in a language for arithmetic, built up using functional symbols. Such terms are semantically complex, because their semantic evaluation is non-trivially compositional. But that semantic complexity does not imply any complexity in the corresponding referents or semantic values, which are (in this case) simply natural numbers. After all, by elementary arithmetic, equations such as ‘7 + 5 = 12’ are *true*, where ‘=’ means strict identity.[6,7]

Philosophers tend to think of all semantically complex singular terms on the model of definite descriptions, and therefore as not *directly referential*, as not contributing only their referent to the compositional semantic evaluation of the larger expressions in which they

occur. But that is a mistake. On the natural semantics for functional expressions just sketched, what they contribute to compositional semantic evaluation is of exactly the same type as what atomic singular terms contribute, their referent—in this case, a number. Despite their semantic complexity, such functional expressions are directly referential. We have learnt from Kripke, Kaplan, and others not to assimilate individual constants to definite descriptions; the underlying lesson applies to semantically complex functional expressions too. In particular, unlike definite descriptions, functional expressions are scopeless, so the inter-substitution of the coreferential directly referential functional expressions '8', '9 – 1', '7 + 1', '*p*(9)', and '*s*(7)' in (4), (5), (6), (5*), and (6*) preserves truth, without risking some sort of scope fallacy. Thus, all those sentences have the same truth-value. Since (4) is trivially true, (5), (6), (5*), and (6*) are true too. That (5) and (5*) are false is an illusion.

The claim is not that the singular terms '8', '9 – 1', '7 + 1', '*p*(9)', and '*s*(7)' are all *synonymous*. They all differ from each other in semantic structure, as represented by a syntactic tree where each node is labelled with the semantic value of the corresponding constituent, so in a natural fine-grained sense they are not synonymous. Rather, the point is just that they all have the same final semantic value, the number 8, which is their input to the compositional semantic evaluation of the sentences in which they occur. For the same reason, the sentences (4), (5), (6), (5*), and (6*) are not synonymous in the fine-grained sense either; nevertheless, since they differ only in those singular terms, they all have the same compositionally derived semantic value, and so the same truth-value.

We can extend the argument to examples much closer to (1), (2), and (3), Fine's case of Socrates and {Socrates}. In effect, the curly brackets '{' and '}' constitute a function symbol with a variable number of places: applied to a list (possibly empty) of singular terms, they result in a semantically complex term whose referent is the set of referents of the singular terms on the list. In particular, for singletons, $\text{ref}_{\underline{a}}(\text{`\{'\^{}}t\text{\^{}`\}'}) = \{\text{ref}_{\underline{a}}(t)\}$ for any singular term t and assignment $\underline{a}$. Just as addition has subtraction as a one-sided inverse, so singleton formation has a one-sided inverse: we can define a function symbol, '| |', by stipulating that $|x| = y$ if $x = \{y\}$ and $|x| = x$ if x is not a singleton set. In particular, |{Socrates}| = Socrates (though, since he is not a singleton, {|Socrates|} = {Socrates} ≠ Socrates). Now consider:

(7) It is essential to Socrates that he is |{Socrates}|.

This feels similar to 'It is essential to Socrates that he is a member of singleton Socrates'. There is a natural temptation to judge (7) false, just like (2). After all, {Socrates} is supposed to be extraneous to Socrates himself. But '|{Socrates}|' is simply a functional expression whose referent is Socrates, so replacing '|{Socrates}|' by 'Socrates' in (7) should preserve its truth-value. But the result of the replacement is just the trivially true (1). Thus, the judgment that (7) is false was itself incorrect.

Again, the point is not that the singular terms 'Socrates' and '|{Socrates}|' are synonymous. They differ in semantic structure and so, in the fine-grained sense, are not synonymous. Rather, the point is just that they have the same final semantic value, the man Socrates, which is their input to the compositional semantic evaluation of the sentences in which they occur. For the same reason, the sentences (1) and (7) are not synonymous in the

fine-grained sense either; nevertheless, since they differ only in those singular terms, they have the same compositionally derived semantic value, and so the same truth-value.

At this point, hyperintensionalists about essence may start trying to construct an alternative semantics for functional expressions, one which avoids treating them as directly referential, and gives them complex enough semantic values to generate differences of truth-value under the relevant replacements—without giving the game away by treating the operator 'It is essential to Socrates that' as creating a merely quotational, metalinguistic context and thereby undermining its pretensions to reveal deep metaphysical structure. Even apart from the obvious risk of overfitting the data, such a reaction does not touch the heart of the problem, because it fails to address the challenge posed by the simple, natural semantics just sketched. We *can* use functional expressions with this straightforward semantics; when we do so, we are tempted to make logically untenable assignments of truth-values highly reminiscent of those on which the alleged counterexamples to modal theories of essence rely. In such cases, our judgments of essence are liable to error; our reliance on them is misplaced.

This argument from functional expressions does not involve any comparison between intensional and hyperintensional semantic frameworks. Possible worlds or the like were not mentioned. No appeal was made to necessary equivalence, only to simple identity, for instance of natural numbers by ordinary mathematical standards. Merely 'going hyperintensional' would be irrelevant to the challenge.

A better response to the cases is to learn from our mistakes, by trying to understand how elementary truths such as (5), (5*), and (7) can look false. Once we have diagnosed the mistake, we can ask whether it is also at work in our assessments of Fine's examples, and in particular (2). That is the business of the next section.

4. Essence, explanation, and irrelevance

Given an iterative conception of natural numbers, what may first strike one as off about (5) and (5*) in relation to the essence of 8 is the very mention of 9: it looks irrelevant. Similarly, given an iterative conception of sets, what may first strike one as off about (7) in relation to the essence of Socrates is the very mention of {Socrates}: it looks irrelevant. Fine's own descriptions of his examples strongly evokes such reactions, for example when he says that it is not essential to Socrates that he be distinct from the Eiffel Tower because 'there is nothing in his nature which connects him in any special way to it' and that 'it is no part of Socrates' essence that there be infinitely many prime numbers or that the abstract world of numbers, sets, or what have you, be just as it is' (5). Correspondingly, what may first strike one as off about (2), like (7), in relation to the essence of Socrates is the very mention of {Socrates}.

Such reactions to the examples are naturally explained by the operation of a crude *relevance filter*, which monitors the complement sentence 'A' in 'It is essential to X that A' for material irrelevant to answering the question 'What is X?', and rejects the whole statement if such material is found. Thus, it rejects (5) and (5*) because the complement has the constituent '9', deemed irrelevant to answering the question 'What is 8?', and it rejects (2) and (7) because the complement has the constituent '{Socrates}', deemed irrelevant to answering the question 'What is Socrates?' That happens even though the constituent occurs

in the complement as part of a more complex expression which in effect 'neutralizes' the offending material. Such a mechanism explains the illusion of differences in truth-value amongst (4), (5), (5*), (6), and (6*), and between (1) and (7). By treating semantic constituents in isolation, the relevance filter is sensitive to differences in semantic structure between complex expressions whose overall semantic value is nevertheless the same. Thus, the relevance filter is fine-grained *because* it is superficial: it stays on the linguistic surface.

Once we have identified such a mechanism at work, we should also lose confidence in the key Finean judgment that (1) and (2) differ in truth-value, even though we lack the straightforward logical reason for holding it to be mistaken that we have in the other cases. The relevance filter will automatically reject (2) but not (1), without any attempt to determine whether the complement sentences 'He is a member of {Socrates}' and 'He is Socrates' express the same proposition, just as it automatically rejects (7) but not (1), without any attempt to determine whether the complement sentences 'He is |{Socrates}|' and 'He is Socrates' express the same proposition (as they do). Similarly, in the arithmetical case, the filter will automatically reject (5) and (5*) but not (4), without any attempt to determine whether the respective complement sentences 'It is 9 – 1', 'It is p(9)', and 'It is 8' express the same proposition (as they do).

The filter does *not* reject (3), (6), and (6*). Thus, with (3), it does not deem 'Socrates' irrelevant to answering the question 'What is {Socrates}?', although it deems '{Socrates}' irrelevant to answering the question 'What is Socrates?', and with (6) and (6*), it does not deem '7' irrelevant to answering the question 'What is 8?', although it deems '9' irrelevant to answering the same question. These asymmetries are sensitive to our understanding of 'What is' questions about people, sets, and natural numbers. There is surely much more to be said about that, but it is not to our present purpose. The aim is not to give a positive metaphysical account of the examples, but rather to identify a specific cognitive mechanism that induces mistakes in our thinking, mistakes which may have widespread repercussions across much of our metaphysical theorizing.

Of course, irrelevance is one thing, falsity another. Ordinarily, we can handle the category 'true, but irrelevant' without too much difficulty. Why should a relevance filter induce us to categorize some statements about essence as *false*, not merely as irrelevant?

Such transitions from alleged irrelevance to alleged falsity tend to occur in broadly *explanatory* contexts, where the aim is to help someone understand something. For instance, you ask 'How did Tom get injured?' and I answer 'He was knocked down by a yellow truck'. My statement is true, but the truck's yellowness was irrelevant to Tom's injury (we may assume). However, when we come to assess the statement 'John got injured because he was hit by a yellow truck', 'because' raises the stakes; we may suspect that, strictly speaking, including the word 'yellow' in the clause after 'because' falsifies the whole statement, by semantically requiring the yellowness to play some role in answering the question, by explaining John's injury, which it does not. The suspicion may or may not be correct; what matters here is that it is naturally felt.

Something similar may be going on in our talk of essences. We may understand a statement of the form 'It is essential to X that A' as requiring 'A' to be a good answer to the question 'What is X?', one which at least partially explains what X is and excludes material irrelevant to the question, so 'It is essential to X that A' is false when 'A' includes

explanatorily irrelevant material. Such connections between essence and explanation may indeed be congenial to Fine's own metaphysics of essence, which is often interpreted as reviving a more Aristotelian form of essentialism, contrasted with Kripke's austerely modal account (Kripke 1980). Aristotle offers a unified, richly interconnected account of essence, explanation, and definition (see Charles 2000 for a book-length treatment). Adding an irrelevant conjunct to a correct definition of something can easily yield an extensionally incorrect definition, and adding an irrelevant conjunct to a good explanation of something can easily yield a worse explanation.

Of course, just saying 'He is Socrates' by itself does not explain much, except perhaps by implication to an audience already well-informed about Socrates, but it could at least form the start of a good explanation. By contrast, starting an explanation with 'He belongs to {Socrates}' would typically introduce unhelpful, pointless clutter. We prefer explanations to begin at the beginning, to start simple, and to build up complexity as needed. A more natural contrast is between 'He is a human' and 'He belongs to the set of humans'. The former could easily figure in a good explanation, while the latter introduces the same kind of unhelpful, pointless clutter as before. Correspondingly, many will be tempted to judge (8) true and (9) false:

(8) It is essential to Socrates that he is a human.

(9) It is essential to Socrates that he belongs to the set of humans.[8]

Nevertheless, on reasonable assumptions about the modal metaphysics of sets, necessarily, he is a human if and only if he belongs to the sets of humans.

Those sketchy remarks indicate how irrelevance can easily be taken as a mark of error in explanatory and essentializing contexts. But that does not fully *vindicate* the relevance filter, since it is still predicted to generate logical errors like those identified above, given the superficial level at which it operates.

At best, such a relevance filter is a useful *heuristic* for testing essentialist claims, quick and easy to use, often correct, but far from infallible. Psychologists have long studied the major role of heuristics in human cognition. The tradition associated with Gerd Gigerenzer characterizes them positively as 'fast and frugal' methods, boundedly rational, efficient, and surprisingly reliable (e.g., Gigerenzer, Hertwig, and Pachur 2011). The tradition associated with Daniel Kahneman characterizes them more negatively as 'cheap and dirty' methods, irrational and only of limited reliability (e.g., Kahneman, Slovic, and Tversky 1982). The two traditions agree that heuristics play a key role in human cognition, and that they are far from perfectly reliable. Such reliance on heuristic methods may be unavoidable for finite agents.

Heuristics do not wear their merely heuristic status on their sleeve. Many heuristics are embedded in our perceptual systems; we are alerted to our reliance on them by perceptual illusions. For example, our visual system uses colour boundaries in the visual field as a heuristic for the edges of three-dimensional physical objects in the environment, with no overt warning that we are relying on a mere heuristic. We may find out that we have been relying on such a heuristic when we discover that we have been deceived by camouflage. In a

similar way, philosophical paradoxes may warn us of our reliance on general but fallible cognitive heuristics (Williamson 2024). When the heuristic works smoothly, the judgments may feel very clearly correct, but that does not guarantee that they *are* correct.

A relevance filter bears the marks of a typical heuristic. It requires only shallow processing, and so can be applied quickly, easily, and unreflectively. It often helps us focus on what matters, but sometimes sends us astray. It carries no distractingly salient cognitive health warning to give us pause, for instance when we use it to assess essentialist or explanatory claims. The relevance filter is *efficient*; no wonder we use it, despite its fallibility.

Unsurprisingly, the danger of heuristic-generated error is not confined to essentialist discourse. Even within metaphysics, explanatory heuristics may play a wider role.

Here is an example. The word 'because' is often regarded as hyperintensional, sensitive to differences between necessary equivalents (Schnieder 2011). Aristotle already observed the apparent asymmetry between pairs like (10) and (11) (where 'because' takes wide scope):

(10) It is true that grass is green because grass is green.

(11) Grass is green because it is true that grass is green.

It is natural to judge (10) true and (11) false: (10) gets the explanatory priority the right way round; (11) gets it the wrong way round. Nevertheless, necessarily, it is true that grass is green if and only if grass is green.

Now consider this pair:

(12) A square with sides 17 metres long has an area of 289 metres^2 because $17^2 = 289$.

(13) A square with sides 17 metres long has an area of 289 metres^2 because $17^2 = 17^2$.

It is natural to judge (12) true and (13) false: in (12), the explanans '$17^2 = 289$' makes the required connection between '17^2' and '289', and is explanatorily relevant; in (13), the explanans '$17^2 = 17^2$' fails to make the connection, and is explanatorily irrelevant. But there is an illusion. For, on the natural semantics for functional expressions above, '17^2', though semantically complex, is nevertheless directly referential. It has the same semantic value as the numeral '289', the number 289, which is their input to the compositional semantic evaluation of sentences in which they occur. Since (12) and (13) differ only in the substitution of '17^2' for '289', they must have the same semantic value and so the same truth-value. Of course, as proposed explanations, sentence (12) is genuinely more helpful than sentence (13), but that turns out to be a matter of how perspicuously it presents things, not of which things it presents. The truth-value of (12) and (13) is insensitive to such presentational differences, except on a reading of 'because' as implicitly metalinguistic, which is not what friends of hyperintensional metaphysics want.

Once we have identified the illusion of hyperintensionality in (12) and (13), we should be very wary of assuming that the apparent hyperintensionality in pairs such as (10)

and (11) is genuine. The difference between 'It is true that grass is green' and 'Grass is green' may turn out to be merely presentational, like that between '17^2' and '289', and to make no difference to their semantic value, their input to the compositional semantic evaluation of sentences in which they occur. Thus, (10) and (11) may turn out to have the same truth-value after all. Elsewhere, I have explored in more depth the heuristic that arguably generates such illusions of hyperintensionality (Williamson 2024, chapter 3). The general mechanism is to use degree of explanatory helpfulness to assess the truth-value of sentences with trigger words such as 'because' or 'essential', which seem to invite such assessment. Since explanatory helpfulness is sensitive to all sorts of presentational differences between words and sentences not registered by their semantic values, which can be presented clearly or obscurely, perspicuously organized or all muddled up, illusions of hyperintensionality inevitably result. Thus, the evidence for hyperintensionality in metaphysics is much weaker than is usually assumed.

Intensionalism has obvious abductive advantages over hyperintensionalism in the simplicity and strength of its theoretical framework. The compensating advantage of hyperintensionalism was supposed to be its better fit with the evidence, mainly in the form of examples, especially the alleged counterexamples to intensionalism. That apparent evidential advantage is now dissolving under scrutiny. Metaphysicians need to reconsider a wide range of examples that have been taken to favour hyperintensionalism, and how they look from an intensionalist perspective.

Within a framework of intensional semantics, one can reinterpret the defining equations of a Boolean algebra as expressing identities between propositions, just as arithmetical equations express identities between natural numbers, in both cases with functional expressions interpreted as above. For example, the equation $p \vee q = q \vee p$ expresses the commutativity of disjunction. Strictly speaking, in a higher-order setting one should treat the variables 'p' and 'q' as occupying sentence position rather than singular term position, and '$=$' in that context as an operator forming a sentence from pairs of sentences rather than pairs of singular terms, but otherwise logically analogous to '=' between singular terms.[9] The semantic treatment of functional expressions such as $p \vee q$ will be correspondingly analogous to that for functional expressions such as $m + n$ in arithmetic.

The defining equations of a Boolean algebra entail equations such as $p = p \vee (p \wedge q)$ and $p \vee \neg p = q \vee \neg q$, where the two sides of the equation differ in which variables occur, and so potentially in 'subject matter'. A hyperintensionalist might see q as relevant to $p \vee (p \wedge q)$ but not to p, and as relevant to $q \vee \neg q$ but not to $p \vee \neg p$. Correspondingly, the hyperintensionalist might take the proposition expressed by the left-hand side of such an equation (on an assignment of mutually independent propositions to the variables p and q) to differ from the proposition expressed by the right-hand side in its grounds or whatever, and so deny the Boolean equations. The obvious risk in such a strategy is of projecting merely presentational differences onto the propositions presented. We have already seen how easy such mistakes are to make, for example with (4), (5), (6), (5*), and (6*), with (1) and (7), and with (12) and (13). A shallow heuristic for relevance will predictably deliver the judgment that the Boolean equations are false, even if they are true.

A salient difference between the case of natural numbers and that of propositions is that, in the former, denying the equations at issue is not a serious option. Arithmetic tells us that $s(7) = 7 + 1 = 8 = p(9) = 9 - 1$, and denying arithmetic would be obviously foolish. By contrast, although propositional Booleanism tells us that $p = p \vee (p \wedge q)$ and $p \vee \neg p = q \vee \neg q$, denying propositional Booleanism is not *obviously* foolish; the theory is less well-established than arithmetic. The initial appearances of the 'hyperintensional' cases are analogous; it is just that arithmetic has more authority than propositional Booleanism to overrule them.

For (1) and (2), not even propositional Booleanism tells us that 'He is Socrates' and 'He belongs to {Socrates}' express the same proposition (where 'he' refers to Socrates both times). Similarly, for (8) and (9), propositional Booleanism does not tell us that 'He is a human' and 'He belongs to the set of humans' express the same proposition. Their equivalence is derived from a (plausible) modal metaphysics of sets. Nevertheless, given our bitter experience with illusions of difference in comparable cases, including the very similar one of (1) and (7), we should be wary of assuming that (1) and (2) differ in truth-value, or that 'He is Socrates' and 'He belongs to {Socrates}' differ in semantic value.

The support for hyperintensionalist metaphysics, and in particular against an intensionalist-modalist account of essence, is looking flimsy and unsafe. Of course, I have not attempted to examine all the considerations that have been brought against intensionalism. I have assessed others elsewhere (Williamson 2024). This paper targets 'Essence and Modality', the flagship of the hyperintensionalist fleet.

5. *Philosophy in the presence of heuristics*

Human cognition, including our philosophical thinking, is riddled with reliance on fallible heuristics. What are the implications for the methodology of metaphysics?

Whatever we can think, we can doubt: perhaps our thought is the output of some heuristic or other in conditions where it is unreliable. But that is just generic sceptical doubt. It applies in particular to sense perception, which is riddled with reliance on fallible heuristics. That a heuristic generates false belief in bad circumstances is compatible with its generating knowledge in good ones. General fallibility does not warrant general scepticism, otherwise natural science should abandon its use of sense perception, with its pervasive reliance on heuristics.

Generic sceptical arguments invoking the cognitive role of heuristics are quite different from the argument in sections 3 and 4 above against Fine's case for hyperintensionalism in 'Essence and Modality'. My argument depends on an analogy between the pattern of judgments involved in Fine's proposed counterexample to intensional accounts of essence and structurally and phenomenologically very similar patterns of judgments elsewhere that nevertheless turn out to be logically inconsistent, combined with the availability of a plausible, natural, and efficient heuristic to explain our attraction to those patterns of judgment, even when they are erroneous. In the background is the heavy theoretical cost of abandoning the intensionalist framework, with all its well-established abductive virtues, in favour of various half-baked hyperintensionalist alternatives with all

their complications and difficulties. Those conditions are far more specific than the mere generic likelihood that some heuristics are at work somewhere in the vicinity, for good or ill.

How can we do better? One obvious moral is that we should be wary of such an example-driven methodology, since it is so vulnerable to errors in our heuristic-generated judgments about cases.

The point here is not the worry associated with the 'negative program' in experimental philosophy, that the judgments might vary with ethnicity or gender. Cognitive heuristics as general as relevance filters might well be humanly universal. Rather, the point is that, even if humanly universal, heuristics are still a source of error.

None of this means that we should stop using examples in philosophy, any more than visual illusions mean that we should do science with our eyes shut. We need to consider potential counterexamples, to keep our philosophical theorizing honest. In any case, the role of heuristics in philosophy is hardly confined to examples: we may well be relying on them in much of our informal general reasoning too, for instance when we reason by analogy. The challenge is to make our overall philosophical methodology more robust, less vulnerable to the errors that will inevitably occur, since we are human. Science at its best has found such ways; the task is not impossible.

A resolution to 'be more careful' by itself is of little help. It may just lead to applying the same old heuristic more carefully and thereby making the same old mistakes, when the heuristic itself is at fault.

One promising strategy is to use a mix of methods. For example, if arguments from the traditional method of hypothetical cases, from formal model-building, and from results of natural science all point in the same direction, that may well be a good direction to go in, and our evidence is typically much stronger than if we had used only one of those three methods. The examples in 'Essence and Modality' would have been harder to oppose had they been predicted by a general hyperintensional framework theory as simple and strong as intensionalism. Of course, that would be far too much to ask of a single lecture-length article, and Fine did take some steps towards developing hyperintensional theories in other work published at the same time. Nevertheless, in retrospect, the haste with which large parts of the metaphysics community came to treat the intensionalist paradigm as refuted by a few questionable observations, in the absence of a properly developed alternative framework, hardly looks like scientific best practice. As Thomas Kuhn emphasized, no paradigm in science is ever anomaly-free (Kuhn 1962). Fine's examples were indeed anomalous for intensionalism when he produced them, but that should have motivated a search for ways of resolving the anomaly, not the premature assumption that no such way was to be found. As explained above, the anomaly can be resolved: it was an explicable observational error.

I do *not* say that the error should have been obvious all along. That would reflect badly on me, since it took me many years to see what had gone wrong (Williamson 2021, 2024). In 'Essence and Modality', Fine shows the profound creativity of a philosopher who pushes a natural human heuristic to its limit in paradox.

Acknowledgements

This paper originates in a short talk to a symposium celebrating the thirtieth anniversary of the publication of ‘Essence and Modality’, organized by Michael Raven for the 2024 APA Pacific Division Meeting in Portland, Oregon. I then gave an expanded version of the talk at Yale. I thank Kit Fine for his detailed response in Portland, audiences at both events for their comments and questions, and Daniel Kodsi for comments on a draft of this paper.

Notes

1 All page references are to Fine 1994.

2 The theoretical framework of 'analytic truth' and 'conceptual connections' are criticized in Williamson 2007. In the present paper, I grant it to Fine simply for the sake of argument. Much of what is said here in terms of 'concepts' could be rephrased in terms of interpreted words and phrases.

3 One can add a redundant extra argument place to the term 'necessary' to match the source (compare $\lambda x(Fx)$ and $\lambda xy(Fx \wedge y = y)$).

4 For general discussion of overfitting in philosophy see Williamson 2024.

5 The terminology refers to Quine's set theory NF, to which Forster 2019 is an introduction. For impure sets like {Socrates}, we can use the system NFU, a weakening of NF to allow individuals, introduced and proved consistent in Jensen 1969.

6 This can all be easily adjusted to a model theory for the language of arithmetic consistent with a structuralist philosophy of arithmetic.

7 In a system of relational types for a higher-order language (Bacon 2024: 41-44), as opposed to the more standard system of functional types, function symbols may have to be defined away in terms of definite descriptions. However, even if such a system of types is in some sense metaphysically fundamental, semantic theories of human languages need not be given in a metalanguage whose terms carve at metaphysically fundamental joints (reference itself is presumably not a metaphysically fundamental relation). Semantics is not metaphysics. A metalanguage with function symbols, such as the one envisaged, is straightforwardly intelligible and adequate for purposes of semantic theorizing about an object-language with function symbols; it suffices for making the case in the text for semantically complex but directly referential functional expressions. Thanks to Harvey Lederman for raising this issue in conversation.

8 On the necessitist metaphysics defended in Williamson 2013, the modal analogues of (8) and (9) would be something like 'Necessarily, if Socrates is concrete, he is a human' and 'Necessarily, if Socrates is concrete, he belongs to the set of humans' respectively.

9 Bacon 2024: 113-116 provides a brief introduction to propositional Booleanism in the setting of higher-order logic. In what follows, for ease of exposition I disrespect the type distinctions built into higher-order logic, for example by talking of 'propositions' rather than quantifying into sentence position. The main

argument could be reworked into more type-theoretically accurate terms if desired.

References

Bacon, Andrew. 2024: *A Philosophical Introduction to Higher-Order Logics*. London: Routledge.

Charles, David. 2000: *Aristotle on Meaning and Essence*. Oxford: Clarendon Press.

Fine, Kit. 1994: 'Essence and Modality: The Second *Philosophical Perspectives* Lecture', *Philosophical Perspectives*, 8: 1-16.

Forster, Thomas. 2019: 'Quine's New Foundations', *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2019/entries/quine-nf/>.

Gigerenzer, Gerd, Ralph Hertwig, and Thorsten Pachur (eds.) 2011: *Heuristics: The Foundations of Adaptive Behavior*. New York: Oxford University Press.

Jensen, R.B. 1969: 'On the Consistency of a Slight(?) Modification of Quine's NF', *Synthese*, 19: 250–263.

Kahneman, Daniel, Paul Slovic, and Amos Tversky (eds.) 1982: *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Kripke, Saul. 1980: *Naming and Necessity*. Oxford: Blackwell.

Kuhn, Thomas. 1962: *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Nolan, Daniel. 2014: 'Hyperintensional metaphysics', *Philosophical Studies*, 171: 149-160.

Schnieder, Benjamin. 2011: 'A logic for "because"', *Review of Symbolic Logic*, 4: 445-465.

Williamson, Timothy. 2007: *The Philosophy of Philosophy*. Oxford: Wiley-Blackwell. Enlarged ed. 2021.

Williamson, Timothy. 2013: *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

Williamson, Timothy. 2021: 'Degrees of freedom: is good philosophy bad science?', *Disputatio*, 13: 73-94.

Williamson, Timothy. 2024: *Overfitting and Heuristics in Philosophy*. New York: Oxford University Press.