Paradoxical Desires*

Ethan Jerzak

December 31, 2018

Abstract

I present a paradoxical combination of desires. I show why it's paradoxical, and consider ways of responding. The paradox saddles us with an unappealing trilemma: either we reject the possibility of the case by placing surprising restrictions on what we can desire, or we deny plausibly constitutive principles linking desires to the conditions under which they are satisfied, or we revise some bit of classical logic. I argue that denying the possibility of the case is unmotivated on any reasonable way of thinking about mental content, and rejecting those desire-satisfaction principles leads to revenge paradoxes. So the best response is a non-classical one, according to which certain desires are neither determinately satisfied nor determinately not satisfied. Thus, theorizing about paradoxical propositional attitudes helps constrain the space of possibilities for adequate solutions to semantic paradoxes more generally.

1 The paradox

There's almost nothing you can't want. A glass of Bordeaux, a degree in astrophysics, world peace ... You name it, and there's probably someone, somewhere, who wants it.

It often happens that our desires involve other people's desires. Jennifer Aniston doesn't just want her deadbeat boyfriend in the 2006 classic "The Break Up" to do the dishes; she wants him to *want* to do the dishes. Conversely, our desires can involve a blind kind of reference to others' desires. I may not particularly care where we go to dinner, but care very much about the enjoyment of my more opinionated dining comrades. Thus I might want simply to go wherever my comrades most want to go. Similarly a parent might want to get their daughter whatever she most wants for her birthday, without having any idea what that is. Desire-directed desires such as these abound.

That isn't to say that they can't get us into trouble. Imagine that my dining

*Forthcoming in *Proceedings of the Aristotelian Society*. Thanks to Wes Holliday, Arc Kocurek, John MacFarlane, Sven Neth, Rachel Rudoph, and Seth Yalcin for very helpful discussions and comments on drafts. Thanks also to audiences in Berkeley, San Diego, London, Barcelona, and Oxford.

comrades turn out to be as unopinionated about where to dine as I am. Then we might end up in the following sort of situation:

My strongest desire: That we dine wherever my dining comrades most want to dine.

My comrades' strongest desire: That we dine wherever *I* most want to dine.

In this case, we've got a problem on our hands: What I want depends on what my comrades want, and vice-versa, and we're stuck. It will be completely unclear where we should dine until some of us change our desires. (My colleagues and I often languish indecisively in such states.)

We are well-advised to avoid desires such as these when they aren't supplemented by other, more basic desires. They are the cause of many indecisive wallowings among domestic partners, friends, and colleagues. But the situation can get far worse than indecisive strife; desire-directed desires can lead to outright paradox.

Here, it seems, is a perfectly possible pair of desires that could be had by two acquaintances, Mal (think 'malevolent') and Ben (think 'benevolent'):

Mal's strongest desire: That Ben doesn't get whatever he most strongly desires.

Ben's strongest desire: That Mal gets whatever she most strongly desires.

Cases like this surely occur less frequently than the cooperative predicament. We require Mal to harbor a particular kind of ill-will towards Ben, who in turn has nothing but benevolent desires towards Mal. We cannot expect Ben to endure such treatment. But endure it he might. Perhaps Ben is a religious man, albeit a hedonistic one, whose life purpose is to bring about the satisfaction of others' desires. Perhaps Mal is a misanthrope, who wants all benevolent men such as Ben to have their sanctimoniously benevolent desires frustrated. Whatever we say about this case, it represents a *prima facie* possible pair of desires. Mal and Ben might have existed.

Those familiar with the Liar paradox (especially the contingent versions presented in Kripke (1975)) will have presaged a paradox. And indeed, one looms. The problem in the case of cooperative desires was that nothing would happen. In order to know what I want, we have to figure out what my dining comrades want; but in order to know what they want, we have to figure out what I want. And so nothing gets done. A regrettable outcome, but a perfectly coherent one.

In the case of desires like Mal's and Ben's, however, we risk encountering far bigger game: true contradictions. When you desire something, it seems, one of two things can happen. You either eventually get what you want, or you fail to. So we can ask in this situation: whose strongest desire is satisfied here? Mal's? Ben's? Both? Neither? A bit of reasoning shows that none of these answers makes any sense.

Here's the proof. Suppose that Mal gets what she most strongly desires. That just means that Ben doesn't get what he most strongly desires. But Ben simply desired that Mal's strongest desire be satisfied. Since Ben doesn't get what he most desires, neither can Mal. But we supposed that Mal *does* eventually get what she most strongly desires. Thus this assumption must be false; it must be that Mal *doesn't* eventually have her strongest desire satisfied.

However, this hypothesis is no better! Mal most strongly desires Ben's strongest desire to be frustrated; so if she doesn't get what she desires, that can only be because Ben gets what he desires. But since Ben just wants Mal to get what *she* desires, the satisfaction of Ben's desire requires the satisfaction of Mal's. Thus, supposing that Mal's strongest desire is frustrated, we can prove that it's satisfied. Contradiction.

Here's another way to see the contradiction: Because of what Mal wants, Mal's desire is satisfied if and only if Ben's isn't ($m \leftrightarrow \neg b$). But because of what Ben wants, Ben's desire is satisfied if and only if Mal's is ($b \leftrightarrow m$). Therefore Ben's desire is satisfied if and only if it isn't satisfied ($b \leftrightarrow \neg b$), a classical contradiction.

It's important to point out that this is a putative contradiction in the world, not merely in the contents of Mal's and Ben's desires. It's a common idea that I can have incompatible desires: I might want to drink a third glass of wine (because it will taste pleasant), and simultaneously want not to drink it (so as to avoid subsequent hangovers).[1] But this case is different: the contradiction concerns whose desire is, in fact, satisfied. Mal's and Ben's desires are each perfectly internally coherent, in the sense that there is a possible state of the world in which each (individually) is satisfied. Mal's strongest desire would be satisfied, for example, in a world in which Ben most strongly desired a beer and failed to get a beer. But in *our* world, in which Mal and Ben have the desires ascribed above, a plain contradiction follows from any classical way of saying whose strongest desire is satisfied.

[1] See Phillips-Brown (2017) for more on these internally inconsistent desire attributions.

2 Responses

What does this paradox show? Two responses naturally suggest themselves:

Response 1: Deny the possibility of the case. We could hold that the case presents merely *prima facie* possible pairs of desires; actually this description misrepresents the underlying psychological reality.

Response 2: Accept the possibility of the case, but try to block the reasoning that leads to contradiction.

Response 2 comes in two kinds:

Response 2a: Deny a principle of desire satisfaction, particularly the principle that, if S most strongly desires that p, then that S's strongest desire is satisfied if and only if p.

Response 2b: Deny a logical principle—most naturally, either bivalence or non-contradiction.

In what follows, I'll argue that Response 1 and 2a aren't right, and so some version of Response 2b must be. I conclude by drawing out some consequences. In particular, cases like this undermine the assumption, prevalent in the literature on the Liar paradox, that paradoxes of self-reference (broadly construed) pose a problem peculiar to language, and in particular for the expressive power of the truth predicate. If cases like the one I've described are possible, then structurally identical paradoxes arise at the level of thought itself, regardless of the expressive resources of the language in which we express those thoughts. This, I'll argue, provides a new reason for adopting a broadly non-classical approach to semantic paradoxes more generally, because solutions that hold onto classical logic by denying the T-schema are not applicable to these non-linguistic versions. But first: can cases like this even arise?

3 Denying the case

The above proof purports to show that the arrangement of the world putatively described above, though seemingly possible, cannot obtain. Mal and Ben can't mutually desire what they seem to desire.

It's one thing simply to stipulate this; it's another to explain why it is so. What, beyond a non-explanatory inconsistency proof, explains *why* Mal and Ben can't

have the desires that they seem for all the world to have?

A proponent of Response 1 has a choice to make here. She can deny the possibility of desires like Mal's and Ben's intrinsically, or relationally. The intrinsic version of Response 1 denies that anyone can *ever* have desires of the forms that Mal's and Ben's seem to have, regardless of who else desires what. The relational version allows that we can sometimes have desires like those, but holds that when certain global features of the world obtain (features concerning who else desires what), we can no longer have those desires. I'll argue that each of these broad kinds of approach faces insuperable problems, and that neither satisfactorily explains why the Mal/Ben case can't obtain.

3.1 Intrinsic case-denying

According to the intrinsic version of Response 1, we can just never have desires like Mal's and Ben's, regardless of which desires other people happen to have. Indeed, to fully block the possibility of paradoxes like the one above, we'd have to ensure that we never really have desires whose content inextricably involves the satisfaction or non-satisfaction of others' desires. Other, more complicated paradoxes loom, and pretty much any local, desire-satisfaction-involving kind of desire can be turned into a paradoxical one with enough case-rejiggering.

The challenge for this way of responding is to offer some story about why we can't have desires like these, even in innocuous cases. As motivated above, desire-involving desires are commonly attributed in everyday life. So someone who wanted to deny these desires intrinsically would have to give a theory of content that provided some satisfactory error theory for these attributions.

What are the prospects for such a theory? Not good, I think. In recent literature on propositional attitudes and mental content, there are two broad ways that people have theorized about beliefs and desires, and what sorts of contents those states have. On some views, like those of Geach (1957), Sellars and Chisholm (1957), Dummett (1974), Davidson (1975), Fodor (1975), and Field (1978), propositional attitudes are essentially language-like in nature. Believing a proposition is standing in a relation to a sentence in what's sometimes called a "language of thought", and this language has many of the features (compositionality, for example) that we associate with ordinary natural languages.

According to an alternate picture, most prominently advocated by Lewis (1979) and Stalnaker (1984), beliefs and desires aren't linguistic in structure. Instead,

these states are conceived of pragmatically, as fundamentally tied less to the assertion of sentences than to the explanation and prediction of rational action. On many views of this kind, propositions are modeled as something less fine-grained than sentences—sets of possible worlds, or centered worlds, or something like that—together with some sort of story about how relations of believing and desiring to these sorts of things can play a role in the explanatory, folk-psychological theory that is their home.

On neither of these ways of thinking about content is there any good reason to think that desires like Mal's and Ben's can't arise intrinsically.

On the language of thought picture, to have a belief or desire is to stand in a certain kind of relation to some sentence-like object. Which sentences in the language of thought can I believe or desire? Any one, proponents of this view generally think, that can be understood or asserted by the person doing the believing or desiring. Now, we can definitely have *some* thoughts about other peoples' desires. There's no motivation for thinking that we can't *believe* that someone's given desire is satisfied or not. We definitely have the concept of desire. But given that we posses this concept, denying the existence of desire-involving desires seems just as unmotivated as denying the existence of truth-predicate-involving sentences. Since we can entertain some sentences in the language of thought which involve the notion of desire, what prevents us from ever standing in the relation of desiring to them?

Denying that we can desire things about desires would be an ad-hoc and unmotivated constraint from the language of thought perspective. *Why*, beyond the fact that it can lead to Mal/Ben-like situations, did we evolve such that certain language-of-thought sentences can go in the belief-box but never in the desire-box? If anything, desire seems less constrained than belief, not more. The language-of-thought proponent who wanted to deny the case intrinsically would need to cook up some sort of story about why we are forbidden from desiring certain language-of-thought sentences that we can certainly believe, and I can't see a promising way to work it out.

Indeed, Whittle (2017) has shown an impossibility result in this vicinity. He's shown that *if* propositions are structured, as the language-of-thought theorist insists, then self-referential (and hence paradoxical) propositions are unavoidable. He shows how a propositional version of the diagonalization lemma can be proved with very minimal assumptions about how propositions would have to be structured, if they are structured at all.[2] So the only real hope for

[2]The diagonalization lemma is the tool used to generate analogues of self-referential sentences

denying the existence of paradoxical contents (of which the Mal/Ben case is a contingent version) lies with an unstructured theory of content.

On broadly Stalnakerian views, there's no commitment to the idea that propositions have a sentence-like structure that corresponds in any straightforward way to that of the sentences we assert. Instead, thoughts are conceived of as those things the attribution of which can explain why rational agents act in certain ways. So we might attribute to someone a desire to stay dry and a belief that it's raining as part of an explanation as to why they grabbed an umbrella on the way out the door. It's this role that desire/belief attributions are supposed to serve, and they are given only as much structure as required to serve that role. Since it's hard to see how, for example, a belief in $p \wedge q$ could contribute anything different to a rational agent's actions than a belief in $q \wedge p$, many have thought that this structure will be something less fine-grained than that of sentences.

This picture is more promising than the language-of-thought one for denying the existence of desire-involving desires, for this view connects the states of believing and desiring less tightly to language. While we can pretty clearly form *sentences* about someone's desiring something, and about about that desire's being satisfied or unsatisfied, the Stalnakerian doesn't assume that this maps in any isomorphic way to the structure of thought. Instead, thought has only the amount of structure required to explain rational action.

Can non-structured theories make a plausible case for denying desire-involving desires? I don't think they can. Even on Stalnaker's own terms, we cannot get away with forbidding these kinds of desires. The reason is that complicated enough profiles of rational, desire-directed actions can *necessitate* these kinds of attributions. Mal might be acting in ways that make the attribution of Ben-desire-involving desires explanatorily indispensable. Imagine that you observe Mal exhibiting the following behavior. Whenever she sees Ben reaching for something, she swats it away. Whenever Ben applies for a job, Mal destroys the application. Whenever Ben says that he wants something, Mal does everything in her power to prevent its being brought about. In all the nearby counterfactual situations in which Ben has certain desires, Mal is lurking, trying to frustrate them.

If that's how Mal is behaving, there doesn't seem to be any way to describe her state of mind other than with a Ben-desire-involving desire. Say that, in the actual world, Ben happens to desire ice cream. In this world, it's true that

in sufficiently rich mathematical languages. See Boolos, Burgess, and Jeffrey (2007) for details.

Mal desires that Ben not have ice cream. But *just* to say this misses out on an important counterfactually robust feature of Mal's mental state. In a nearby situation in which Ben starts wanting macaroons instead, the theorist who attributed to Mal a desire that Ben not have ice cream won't make as good of predictions as those who attributed to Mal a desire that Ben not get what he most strongly desires. So, even on a Stalnakerian picture, we shouldn't think that desire-satisfaction-involving desires never occur. The only hope for denying the Mal/Ben case is to say that something goes wrong, not with Mal's desire on its own or Ben's desire on its own, but rather with the paradoxical relation in which they happen to stand.

3.2 Relational case-denying

What about the relational version of Response 1? According to this kind of view, it's possible for Ben to have his desire, and it's possible for Mal to have her desire. But what's *not* possible is for Mal and Ben to have these desires together. Let's fix Mal's desire, and say that, initially, their desires are these:

Mal's strongest desire: That Ben doesn't get what Ben most strongly desires.

Ben's strongest desire: That Monika get the job she applied for.

The relational proponent of Response 1 need not deny that this is a perfectly possible situation. The result is that Mal winds up with a desire that (given what Ben desires) is satisfied just in case Monika does not get the job she applied for.

Now suppose that, in this situation, Ben forgets all about Monika, and reflects upon his positive feelings towards Mal. "What a fine person Mal is!" Ben thinks; "I really just hope that she gets whatever it is she most wants."

The question for the relational case-denier is, what should we say about what happens to Mal? The answer, in any case, is going to be strange. Either Ben's forming this desire immediately robs Mal of her previous desire, without changing anything about her, or it prevents Ben from getting into the state of desire he'd otherwise have gotten into if Mal's strongest desire had simply been to drink a nice glass of wine. (Or both.)

Is any of these options plausible? Any of these options involves a commitment to a pretty radical kind of content externalism—too radical, I think, to stomach. Since Putnam (1981) it's been a commonly held idea that which propositions I believe and desire supervenes on more than my own internal psychological

state; it depends partly on features of the world. So here on Earth, I have beliefs and desires involving H2O, but on Twin Earth that exact same internal psychological state would put me into belief and desire relations to XYZ instead. In a similar vein, Burge (1979) argues that the content of my beliefs about arthritis is fixed by more than just my own private understanding of how arthritis works; instead, some such facts (like whether it can occur in my thigh) are fixed by relevant experts. So my belief that I have arthritis in my thigh is false, even though it might be true according to my own internal conceptual scheme.

According to the relational case-denier, something similar is happening here. A single intrinsic psychological state can, depending on the rest of the world, put Mal into different desiderative states. If Ben has normal desires, it puts Mal into a Ben-desire-involving desiderative state; if Ben is in the state described in the original case, this psychological state puts Mal into... well, some other kind of state. (It would be a burden on this kind of theorist to say exactly what Mal's desiderative state is in this case.)

Though externalism about mental content is popular these days—indeed, I count myself as an externalist, largely for Putnam-Burge style reasons—I don't think we should be happy with *this* kind of externalism. Why not? Well, the traditional defenders of content externalism do not simply assert that content supervenes widely, and leave things at that. All such cases come together with a very natural explanation for *why* and *how* our beliefs/desires can differ in content without our differing in intrinsic psychological state. In the case of Twin Earth, Putnam doesn't simply assert that the content of water-like beliefs differs; he gives an account that explains why this is so, and why it makes sense to use a notion of content that behaves that way. The reason why I can't have thoughts about XYZ, but can have thoughts about H2O, is that I've never been in the right kind of causal contact with XYZ to have thoughts about it. It's this causal story of how representational states get their content that explains the wide supervenience in Twin Earth cases.

Similarly for Burge. We have a notion of content that allows for deference to experts because we live in a social community that benefits from a division of intellectual labor regarding the understanding of various phenomena. Doctors get paid to know about arthritis and where it can occur, and I defer to them conceptually because they know about the phenomenon in question better than I do. Their expertise, and my deference to it, naturally explains why and how content supervenes widely in the way that it does.

It's really hard to see what a similar kind of explanation would look like in

the Mal/Ben case. Mal and Ben appear to have all the concepts necessary to entertain the thoughts they seem to be entertaining. They are perfectly acquainted with each other, and both of them have the concept of desire. Both of them know what it is for a desire to be satisfied or not. So they have all of the concepts and acquaintances with the right objects in order to have the thoughts/desires they seem to have. The *only* reason to think that they can't have those desires in this global case is that they are (classically) mutually inconsistent; we have as yet no theory that explains why the content of our desires can depend on the whims of other people in this particular way. And, unlike in the case of H2O and XYZ, we have no good story about how even to describe Mal's and Ben's desiderative states instead.

So it's better to allow that Ben and Mal can have these desires, mutually, in this case. That only leaves the pesky little problem of the paradox. What's going on, if we seem to have a proof that this situation, which we have no good independent reason to think can't arise, can't arise?

4 Denying a principle of desire-satisfaction

The only hope for retaining classical logic while holding onto the possibility of the Mal/Ben case lies in denying a non-logical principle that goes into the derivation of a contradiction. Let us therefore look more closely at how this contradiction was derived.

One way was via the biconditionals $m \leftrightarrow \neg b$ and $b \leftrightarrow m$. These biconditionals each have intuitive plausibility, even in the paradoxical case. But this intuitive plausibility can be bolstered; they are derivable from seemingly incontrovertible principles relating desires to their conditions of satisfaction. The thought is that Mal's state of desire gives rise to, and indeed is characterized by, the first biconditional. What Mal wants is for Ben's desire to be frustrated; to give her what she wants is just to frustrate Ben's desire, and to fail to give her what she wants is just to satisfy Ben's desire. So her desire is satisfied just in case Ben's isn't. Lying behind this thought is a principle of desire satisfaction:

> **Desire satisfaction schema (DSS)**: If S desires that p, then S's desire that p is satisfied if and only if p.

The relevant instance is:

> If Mal desires that Ben's strongest desire be frustrated, then her desire that Ben's strongest desire be frustrated is satisfied if and

only if Ben's strongest desire is frustrated.

This principle suffices to generate, in classical logic applied to the Mal/Ben case, the biconditional $m \leftrightarrow \neg b$. For Mal's strongest desire *is* that Ben's strongest desire fail to obtain. Thus, by (DSS), it is satisfied if and only if Ben's strongest desire is frustrated.

(DSS) has intuitive plausibility. But there is good reason to doubt that it is always true. Desire reports are generally thought to contain normality presuppositions as part of their meaning. So, imagine that John wants to drink a beer, and you give him a poisoned beer. He drinks it and dies. Here, it seems like it's true that John drinks a beer; and it's true that John desired to drink a beer; but plausibly, John does not get what he wants.

You might wonder if this has something to do with the indefinite article: John's real desire is to get a [normal, non-poisoned] beer. But this behavior seems to persist even without indefinite articles. Suppose John forms a desire to drink *this* beer in front of him, not knowing that his enemies have poisoned it. (DSS) commits us to:

John's desire to drink this beer is satisfied iff John drinks this beer.

Here, intuitions may not be as clear. It seems like there is *some* sense in which, if John drinks this beer, his desire is satisfied. It's just not satisfied in the way he wanted it to be satisfied—the normal, refreshing, non-poisoned way. This way of describing things seems relatively natural, and it does not involve denying (DSS). John's desire to drink this beer is satisfied, just not how he would have liked.

But there is another temptation to say that John doesn't actually get what he wants. After all, drinking the poisoned beer is really bad for John; if John knew that the beer were poisoned, he would immediately renounce his desire for it, and would have *no* inclination to drink it. So it's odd to say that he gets what he wants in spite of his untimely and undesired demise.[3]

I feel the pull of both intuitions. Thus, I won't rely on (DSS) in my argument

[3]The kind of desire report at play behind this intuition is what I've called the **advisory** kind in Jerzak (n.d.). Attributers of such desire reports can help themselves to some of their information (of which the desirer herself may be ignorant) in working out what would put the desirer into preferred states, and attribute the desire on that basis. (Said of someone with no knowledge of the MTA: "Sally is heading to Harlem? She doesn't know it, but she wants to take the A train".) The paradox here is better understood as involving the more familiar **predictive** kind of desire attribution, because it should be no surprise that advisory uses are radically externalist in the sense of §3.2.

against this way of blocking the paradox. For the desire satisfaction principle *restricted to strongest desires* is much more plausible. That principle says:

> **Strongest desire satisfaction schema (SDSS)**: If S most strongly desires that p, then S's strongest desire that p is satisfied if and only if p.

This schema is not subject to the objections to the fully general (DSS). If John claims most strongly to desire a beer, receives a poisoned one, and protests that he didn't get what he wanted, it would be reasonable to retort that he must not have *most strongly* desired a beer after all.[4] His subsequent protestation shows that he had another, logically stronger desire all along: a desire for a non-poisoned beer. Desires may be the sorts of things that can fail to be satisfied, even while their content comes true. It may, after all, come true in deviant ways—ways inconsistent with stronger desires that the agent has. But strongest desires aren't like that. Dissatisfaction with deviant ways of satisfying them merely show that they weren't actually strongest desires after all.

Perhaps there are lingering doubts about this. Perhaps even logically strongest desires have deviant ways of failing to be satisfied, such that it could make sense to say that someone most strongly desired p, and p was indeed the case, but they failed to get what they most strongly desired. Even if so, related paradoxes loom for which such a response wouldn't apply. For nothing in the derivation of the contradiction depends very much on the notions of desire and satisfaction in particular. We could have formulated the paradox in terms of closely related attitudes and relations, such as *seeming* desire, or desires *seemingly* being satisfied, or seeming desires seemingly being satisfied, etc. The weaker these attitudes and relations get, the less plausible becomes the story of why the analogue of (SDSS) should fail; but the formal derivation of the paradox isn't affected thereby.[5]

[4] There's an ambiguity here in the phrase "strongest desire" that hasn't mattered much until now. On one reading, it means a desire that is strongest in terms of *content*; that is, a desire whose content entails that of all the agent's other desires. On another reading, it means something like "most psychologically salient/forceful desire", which may fail to be logically strongest. It's the former reading we need for these purposes.

[5] The situation here is similar to revenge versions of Prior's paradox discussed by Prior (1961), Bacon, Hawthorne, and Uzquiano (2016), Bacon and Uzquiano (2018), and Bacon (n.d.). Our desire paradox, suitably formulated in a language with propositional quantifiers and a propositional definite description operator, can be seen as a contingent version of Prior's paradox. The contingency makes particularly vivid the costs of classical propositional logic in this context, for, unlike intrinsically paradoxical versions (in which, say, you believe that all of your beliefs are false or desire that all your desires be frustrated), there doesn't seem to be anything particularly irrational or unrealistic about Mal's and Ben's desires considered

(SDSS), applied to the Mal/Ben case, is all that we need to generate the biconditionals $m \leftrightarrow \neg b$ and $b \leftrightarrow m$. And classical logic is all that is needed to generate contradictions (and triviality) from these biconditionals.

I'll argue that certain principles of classical logic are indeed the root of the paradox, and that giving them up yields a *unified* solution to all paradoxes with this structure. I'll start making this case by drawing out analogies and disanalogies to the more traditional Liar paradox.

5 Non-classical thought

Readers familiar with semantic paradoxes will recall Kripke (1975)'s contingent paradoxes of self-reference. Mal's and Ben's desires are reminiscent of the following pair of sentences:

> The sentence just below this one isn't true.
>
> The sentence just above this one is true.

Is the upper sentence true, or not? If it's true, then the sentence below it isn't, which means that it can't be true either. Thus it must not be true. But then the sentence below it *is* true, in which case the upper sentence must be true after all. The reasoning that leads to paradox is similar to that at play in our desire paradox.

The *matter* of the paradoxes, however, is quite different. Self-referential sentences, like the Liar and Curry sentences, have been said to give rise to *semantic* paradoxes. It's commonly thought that these paradoxes stem, at bottom, from languages too expressive for their own good. Here, for example, is Ramsey explaining why we need to keep an object-language truth predicate at the cost of facing paradoxes:

> We get statements from which we cannot in ordinary language eliminate the words 'true' or 'false'. Thus if I say 'He is always right', I mean that the propositions he asserts are always true, and there does not seem to be any way of *expressing* this without using the word 'true'. (Ramsey 1927, my emphasis)

individually. See also Caie (2012) for similar paradoxes involving belief, where he argues that the rational response to paradoxes like the Liar is to adopt indeterminate states of belief, according to which you neither believe nor fail to believe the Liar sentence.

And here is Field explaining his view of the dialectic concerning paradoxes and classical logic:

> There is little reason to doubt the correctness of classical logic as applied to our most serious discourse, e.g. our most serious physical theories. But the semantic paradoxes arise because truth *talk* gives rise to some anomalous applications (e.g. "viciously self-referential" ones), and it's rash to assume that classical logic continues to be appropriate to these applications. (Field 2016, my emphasis)

Common to both of these passages is what I'll call a *linguistic* diagnosis of these paradoxes. Paradoxes arise, according to this diagnosis, because we want to express certain things (like infinite conjunctions—see Picollo and Schindler (2017)) that we cannot express in a finite way without a truth predicate that behaves disquotationally. This need for expressive power yields a tool that is, in some sense, too powerful for classical logic. The lesson is that a naïve truth predicate and classical connectives are two expressive tools that cannot be combined without triviality. It's because of the way we need to talk that paradoxes of self-reference arise.

The traditional menu of solutions developed in response is similarly informed by this linguistic conception of how the paradoxes get going. Their lesson, according to orthodoxy, is something like this: When developing a formal language, you have to be careful to avoid certain natural combinations of expressive tools that, when combined, yield triviality. Either you must avoid saddling a language with its own truth predicate (as Tarski and Russell urge). Or, if you do insist on having an object-language truth predicate, you've got to weaken it in some unnatural way—either by having sentences with semantic value 1 that aren't true, or by having sentences with semantic value 0 that are true. Or you've got to weaken the logic to something less powerful than classical logic. Perhaps Kleene's K_3; perhaps Priest's Logic of Paradox; perhaps something else. The point is, these paradoxes are thought to have something fundamentally to do with language and expressive power. Thus the label "semantic".

Our desire paradox, despite involving structurally similar reasoning, doesn't have anything to do with language in particular. It doesn't make use of any particularly semantic properties like truth or falsity. Instead, it's just about people getting what they want, or failing to. Of course, we *use* language to *talk* about people getting what they want; but when we do so, we aren't talking about language, as we're obviously doing when we attribute truth or falsity to

sentences as in the traditional semantic paradoxes.

Indeed, the ability of Mal and Ben to speak any particular language is inessential to the paradox. Mal and Ben needn't have any thoughts about which sentences in which languages are true or false in order to desire what they desire. Each just needs to have attitudes about the other's desires, regardless of which language they are expressed in. Regardless, indeed, of whether they are expressed in *any* language.

To make the point particularly strongly, it may even be possible—although I do not stake much on this claim—for sufficiently sophisticated non-linguistic creatures to get themselves into desiderative situations similar to Mal's and Ben's. Desires are things that pretty much any sentient being can have. Horses desire to roam free, tigers to hunt their prey, antelope to flee their predators, all without speaking a language to communicate these desires. These non-linguistic creatures also have the capacity to form other-directed desires: A mother fox might desire the satisfaction of her offspring's desires, or a malicious cat the frustration of his owner's.[6] Thus we might as well imagine that Mal and Ben are two cats, who cannot speak but can be respectively malicious and benevolent. If that's right, it would make a particularly strong case that paradoxes with a Liar-like structure arise at the level of thought itself.

We know from the literature on contingent Liar sentences that the way to solve those paradoxes isn't to deny the existence of the problematic sentences. Paradoxical sentences can clearly be formulated in natural language, and a minimal amount of mathematics make them unavoidable in formal ones. I've argued above that this is so for paradoxical propositions as well: we should not deal with the problems they raise by denying that they exist. Instead we have to deal with them by modifying something else in our theories.

What kind of modification should that be? I'll conclude by suggesting that this paradox gives us a new abductive reason to prefer broadly non-classical treatments.

Solutions to the semantic paradoxes come in two flavors: those that keep classical logic, and those that revise some part of it. Those that keep classical logic must deny one of the T-Schemas:

$$T(\ulcorner s \urcorner) \rightarrow s \qquad \text{(T-out)}$$

[6] I have met such cats.

$$s \rightarrow T(\ulcorner s \urcorner) \qquad \text{(T-in)}$$

The ability to retain these intuitive schemas unrestrictedly has long been touted (for example by Priest (1987) and Field (2008)) as a mark in favor of non-classical approaches. Most of the work that goes into classical approaches, therefore, is concerned with retaining something as close to these schemas as possible without reinviting paradox.

The non-logical (SDSS) is the clear analogue to the T-schema. Recall that this principle says:

> **Strongest desire satisfaction schema (SDSS)**: If S most strongly desires that p, then S's strongest desire that p is satisfied if and only if p.

Defenders of classical logic will point to this principle as the culprit in the desire case, just as they point to the T-schema as the culprit in the Liar paradox.

However, the T-Schema and (SDSS) are about different things. The T-schema relates syntactic objects (sentences) to their truth conditions; (SDSS) relates desires to their satisfaction conditions. This makes a difference. For example, shifts in facts about meaning/conventions can change which sentences are true, but they cannot change what I desire (unless I desire something about meanings or conventions).

To see this, let us grant that "Beer is red" is false, and that I currently desire red wine, though not beer. Consider the following counterfactuals:

1. If 'beer' had meant what 'red wine' means, 'Beer is red' would be true.
2. If 'beer' had meant what 'red wine' means, I would want a beer.

Plausibly, (1) is true, while (2) is false. Shifting the meaning of 'beer' changes the truth conditions of sentences, but not the satisfaction conditions of my wine-related desires. Thus, (SDSS) and the T-Schema are not just trivial notational variations. They have substantively different content, and different modal profiles. That means that anyone who wants to block the desire paradox by denying (SDSS) does not get that move for free, by re-telling whatever story she told to deny the T-Schema. A different kind of story would have to be told in favor of giving up each one.

A non-classical approach, on the other hand, can solve the Liar paradox and the desire paradox by abandoning *the exact same principles.* Both paracomplete and paraconsistent approaches are possible. I'll conclude by briefly sketching

a paracomplete story along the lines of Field (2008), not because I think it's inevitable, but because I favor it over paraconsistent approaches for a broad class of intensional semantic paradoxes (for example the Knower paradox—see Jerzak (2017) for a more involved technical exposition). I won't delve into the technical details here; instead, I'll show how it can yield an attractive package of results in the Mal/Ben case.

In a nutshell, a theory of this kind holds that it's *indeterminate* who gets what they want in the Mal/Ben case. The claim that Mal gets what she wants falls into a truth-value gap, as does its negation. The claims that Mal gets what she wants if and only if Ben doesn't, and that Ben gets what he wants if and only if Mal does, are true.[7] The classical argument from those biconditionals to explosion fails at the negation-introduction step: indeterminate claims are not such that we can suppose them, derive a contradiction, and infer their negations. When possibly indeterminate sentences/propositions are involved, we must be careful not to reason as if they have classical truth values.

This theory denies only two classical rules of inference, both involving suppositional reasoning: negation-introduction, and if-introduction.[8] A unifying virtue of this theory, in our context, is that it attributes the error in reasoning in the Liar paradox, and in the desire paradox, to the exact same steps—in this case, the negation-introduction step. Classical theories, on the other hand, must attribute the error to two different kinds of satisfaction/truth principles—indeed, both of which enjoy immense intuitive support. Thus the classical theorist sees disunity where we ought to have expected unity. The ability to solve two structurally similar paradoxes in the exact same move is a mark in favor of non-classical approaches.

Such a theory still comes with a certain kind of surprising externalism. Facts about whether my desires are satisfied can have non-classical semantic values, and *that* will supervene on more than just my local situation. For instance, Mal's strongest desire would have a classical truth value if Ben's strongest desire were for a beer, and becomes indeterminate the moment Ben forms the desire ascribed above. However, we should sleep easier with this kind of externalism than with the radical externalism that the relational case-denier must espouse. It's a common idea that our beliefs and desires go from true to false, or from satisfied to unsatisfied, according to the whims of the world. On

[7] These biconditionals have to be formulated with a more complicated, non-truth functional conditional, instead of the material conditional. But this conditional collapses into the material conditional in bivalent contexts, so not much is lost by this.

[8] Negation-introduction: $\varphi \vdash \bot \implies \vdash \neg\varphi$; if-introduction: $\varphi \vdash \psi \implies \vdash \varphi \to \psi$.

this paracomplete view, they can go from determinately satisfied/unsatisfied to indeterminate just as easily. However, facts about *what* I desire are more well-behaved. They may supervene widely, but only for the tractable, familiar reasons explored by Burge and Putnam.

This is a more attractive package of views, I think, than the classical alternatives. It all amounts to a new consideration in favor of non-classical approaches to paradoxes in the family of the Liar. Not only must classical approaches invalidate extremely plausible inferential principles involving truth (T-out and T-in), they must also either forbid certain seemingly possible combinations of desires, or else the independently plausible (SDSS). The non-classical approach I outlined solves both sentential and non-sentential kinds of paradoxes in the exact same move. This is a new mark in its favor.

Bibliography

Bacon, Andrew. n.d. "Radical Anti-Disquotationalism." *Philosophical Perspectives*.

Bacon, Andrew, John Hawthorne, and Gabriel Uzquiano. 2016. "Higher-Order Free Logic and the Prior-Kaplan Paradox." *Canadian Journal of Philosophy* 46 (4-5): 493–541.

Bacon, Andrew, and Gabriel Uzquiano. 2018. "Some Results on the Limits of Thought." *Journal of Philosophical Logic*, 1–9.

Boolos, George S., John P. Burgess, and Richard C. Jeffrey. 2007. *Computability and Logic*. 5th. Cambridge: Cambridge University Press.

Burge, Tyler. 1979. "Individualism and the Mental." In *Midwest Studies in Philosophy Iv: Studies in Metaphysics*, edited by Peter French, Theodore E. Uehling, Jr., and Howard K. Wettstein, 73–121. Minneapolis, MN: University of Minnesota Press.

Caie, Michael. 2012. "Belief and Indeterminacy." *Philosophical Review* 121 (1): 1–54.

Davidson, Donald. 1975. "Thought and Talk." In *Mind and Language*, edited by Samuel D. Guttenplan, 1975–7. Clarendon Press.

Dummett, Michael. 1974. "Frege: Philosophy of Language." *Philosophical Quarterly* 24 (97): 349–59.

Field, Hartry. 1978. “Mental Representation.” *Erkenntnis* 13 (July): 9–61.

———. 2008. *Saving Truth from Paradox.* Oxford University Press.

———. 2016. “Indicative Conditionals, Restricted Quantifiers and Naive Truth.” *Review of Symbolic Logic*, 1–28.

Fodor, Jerry A. 1975. *The Language of Thought.* Harvard University Press.

Geach, Peter. 1957. *Mental Acts.* Routledge; Kegan Paul.

Jerzak, Ethan. 2017. “Non-Classical Knowledge.” *Philosophy and Phenomenological Research.*

———. n.d. “Two Ways to Want?” *Journal of Philosophy.*

Kripke, Saul. 1975. “Outline of a Theory of Truth.” *Journal of Philosophy* 72: 690–716.

Lewis, David. 1979. “Attitudes *de Dicto* and *de Se*.” *Philosophical Review* 88: 513–43.

Phillips-Brown, Milo. 2017. “I Want to, but....” *Proceedings of Sinn Und Bedeutung* 21.

Picollo, Lavinia, and Thomas Schindler. 2017. “Disquotation and Infinite Conjunctions.” *Erkenntnis*, July. https://doi.org/10.1007/s10670-017-9919-x.

Priest, Graham. 1987. *In Contradiction: A Study of the Transconsistent.* Dordrecht: Martinus Nijhoff.

Prior, A. N. 1961. “On a Family of Paradoxes.” *Notre Dame Journal of Formal Logic* 2 (1): 16–32.

Putnam, Hilary. 1981. *Reason, Truth and History.* Cambridge: Cambridge University Press.

Ramsey, F. P. 1927. “Facts and Propositions.” *Proceedings of the Aristotelian Society* 7 (1): 153–70.

Sellars, Wilfrid S., and Roderick M. Chisholm. 1957. “Intentionality and the Mental: A Correspondence.” *Minnesota Studies in the Philosophy of Science* 2: 507–39.

Stalnaker, Robert. 1984. *Inquiry.* Cambridge, MA: MIT Press.

Whittle, Bruno. 2017. “Self-Referential Propositions.” *Synthese* 194 (12): 5023–37.