Counterfactual donkeys don't get high[1]

Michael DEIGAN — *Yale University*

Abstract. I present data that suggest the universal entailments of counterfactual donkey sentences aren't as universal as some have claimed. I argue that this favors the strategy of attributing these entailments to a special property of the similarity ordering on worlds provided by some contexts, rather than to a semantically encoded sensitivity to assignment.

Keywords: donkey sentences, counterfactuals, conditionals, similarity, simplification.

Many indicative donkey sentences have universal entailments by default. From the truth of an utterance of (1), for instance, we can normally infer the truth of sentences like those in (2).

(1) If Balaam owns a donkey, he beats it.

(2) a. If Herbert is a donkey Balaam owns, Balaam beats Herbert.
    b. If Eeyore is a donkey Balaam owns, Balaam beats Eeyore.
    c. If Platero is a donkey Balaam owns, Balaam beats Platero.

The natural way to account for this is to give a semantics for indefinites, pronouns, and indicative conditionals on which (1) has a reading semantically equivalent to (3).

(3) $\forall x((\texttt{donkey}(x) \wedge \texttt{Balaam-owns}(x)) \rightarrow \texttt{Balaam-beats}(x))$

This paper is about *counterfactual* donkey sentences, like (4).

(4) If Balaam owned a donkey, he would beat it.

Like their indicative counterparts, such sentences seem to have universal entailments. The truth of (4), for instance, seems to entail the following:

(5) a. If Herbert were a donkey Balaam owned, Balaam would beat Herbert.
    b. If Eeyore were a donkey Balaam owned, Balaam would beat Eeyore.
    c. If Platero were a donkey Balaam owned, Balaam would beat Platero.

The natural way to account for this is to give a semantics for indefinites, pronouns, and counterfactual conditionals on which (4) has a reading—a 'high' reading—which is semantically equivalent to (6).[2]

(6) $\forall x((\texttt{donkey}(x) \wedge \texttt{Balaam-owns}(x)) \mathrel{\square\!\!\rightarrow} \texttt{Balaam-beats}(x))$

[1]For helpful discussions of this material, thanks to Kyle Blumberg, Sam Carter, Simon Goldstein, Maribel Romero, Zoltán Gendler Szabó, Nadine Theiler, Tim Williamson, and especially Lucas Champollion, as well as to four anonymous reviewers, audiences at SuB22 and the Yale Semantics Reading Group, and the particpants of Lucas Champollion's Spring 2017 seminar at NYU on Counterfactuals and Inquisitive Semantics.

[2]Van Rooij (2006) is the first to pursue this strategy, as far as I know, though it's similar to various attempts to give a semantics which validates simplification of disjunctive antecedents. The terminology is from Walker and Romero (2015), who follow van Rooij in this approach.

Proceedings of Sinn und Bedeutung 22, vol. 1, ZASPiL 60, pp. 367–384. ZAS, Berlin.

I will argue against this approach and defend the view that these entailments arise as a byproduct of a special kind of similarity ordering on worlds which the counterfactual conditional takes as an input from context. Counterfactual donkey sentences don't get high readings, but only appear to because in many contexts, the default similarity ordering is of the relevant special kind.

We will proceed as follows. In §1, I lay out the relevant background, including the semantics to be defended (based on Wang (2009)), the criticism of it made by Walker and Romero (henceforth 'WR'), and the competing kind of semantics van Rooij offers that gives counterfactual donkey sentences a high reading. In §2, I rebut WR's criticism. I argue that even the amended semantics requires the kind of special ordering which they wish to reject. This undermines WR's argument in favor of the competing account and against the original proposal. Further, I propose an account of the similarity relation which predicts the special ordering. Its tenability shows WR's argument to be unpersuasive. Finally, in §3, I criticize the accounts that *do* allow high readings. Such accounts, I claim, make incorrect predictions in cases where the antecedent is actually satisfied—and only satisfied by things for which satisfy the consequent—but where the universal entailments do not hold for merely possible antecedent satisfiers. Intuitively, such sentences have no false readings, but on a high reading they would be false. Non-high-reading accounts which get high entailments through special ordering relations make the correct prediction in these cases. High reading accounts do not.

1. Background

There are various approaches to counterfactuals, indefinites, and donkey pronouns which we might try out in dealing with counterfactual donkey sentences. I will limit the current discussion to the ordering semantics approach to counterfactuals developed by Stalnaker and Lewis[3] and dynamic binding approach to indefinites and donkey pronouns based on ideas in Groenendijk and Stokhof (1991) and Groenendijk et al. (1996).[4] This is primarily because the extant explicit discussions of counterfactual donkey sentences in the literature—van Rooij (2006); Wang (2009); Walker and Romero (2015)[5]—all use such theories.[6] I would expect most of the central points I will make to carry over to other frameworks, but will not explore this here.

Let's start with a review of ordering semantics. The basic components are a set of possible worlds W and for each world w a 'similarity' relation $\leq_w$ over $W \times W$. Intuitively, $w_1 \leq_{w_0} w_2$ means that w_1 is more similar to w_0 than w_2 is (in a contextually relevant and currently theoretically underdetermined sense of 'similar') We'll treat this set of similarity orderings as an element $\mathcal{S}$ of the model, which varies depending on the context of utterance.

[3]See, e.g., Stalnaker (1968) and Lewis (1973).

[4]As well as Heim (1982) and Kamp (1981), though less directly.

[5]Though Walker and Romero ultimately propose, for reasons orthogonal to those discussed here, that we move to a dynamic-strict account of counterfactuals, along the lines of von Fintel (2001). Except for the occasional footnote, I will ignore this revision. Most of their discussion proceeds independently of it.

[6]Other recent work on counterfactual donkey sentences includes Walker (2017) and Carter and Goldstein (ms), and came to my attention after I had written this paper. I hope to address these accounts in future work.

As we'll see, what determines the similarity orderings, given a context, will differ from theory to theory. But like most semanticists using this framework, we will hold that it induces a partial order on W—that it is reflexive, transitive, and anti-symmetric. Further, we will assume that it is strongly centered—that for all w, w', if $w \neq w'$, $w <_w w'$; each world is strictly closest to itself. And for ease of exposition, we will assume that similarity relations satisfy the limit assumption, which says that for any world w and non-empty proposition P, there is a w' such that $w' \in P$ and $w' \leq_w w''$ for any $w'' \in P$.[7]

We give the semantics for a counterfactual conditional in two stages, first defining a selection function f, which given a sentence and a world will return the worlds closest to the given world at which the given sentence is true. Given similarity relations and an interpretation function $[\![\cdot]\!]$ which returns a set of worlds for any sentence (namely, those worlds at which the sentence is true in the relevant model),

(7) $\quad f(A,w) = \{w' : w \in [\![A]\!] \wedge \neg\exists w''(w'' \in [\![A]\!] \wedge w'' <_w w')\}.$

Now we can state the truth conditions of a counterfactual conditional as follows:

(8) $\quad [\![A \mathrel{\square\!\!\rightarrow} C]\!] = \{w : \forall w'(w' \in f(A,w) \supset w \in [\![C]\!])\}$

In other words: $A \mathrel{\square\!\!\rightarrow} C$ is true at a world when all the worlds closest to it (according to the similarity relation of the context of utterance) at which A is true are worlds at which C is true.

Clearly, the meaning assigned to counterfactuals by this theory is heavily dependent on what determines the similarity orderings for a given context. To make substantive predictions about the truth-conditions of an utterance of a counterfactual, some details of how these orderings get determined must be provided.

For the moment we will follow the authors under discussion (van Rooij, Wang, WR) in relying on an intuitive notion of similarity in all respects. But throughout we should keep in mind the fact it is well known that this approach is inadequate. For the ordering semantics to get right cases like the famous one from Fine (1975), an intuitive notion of similarity will have to be replaced by something else.

(9) a. If Nixon had pressed the button, there would have been a nuclear holocaust.
b. If Nixon had pressed the button, the wire would miraculously malfunction.

It's easy to imagine scenarios where we want (9a) to come out true rather than (9b), even though a world where the button was pressed but the world was saved by a miraculous wire malfunction would be intuitively more similar to the evaluation world.

For a view on similarity which accounts for our judgments about this case and is closer to being tenable overall, I refer the reader to Lewis (1979), who proposes a system of weighted factors

[7]For discussion, see Lewis (1973: §§1.3–1.4, 2.3, 2.7). Strong centering for the ordering of worlds, it's worth noting, is assumed by van Rooij, Wang, and WR.

that normally go into determining similarity. Roughly, the idea is that what's most important to determining similarity is the amount of widespread violation of physical laws (the fewer miracles, the better). Overall similarity in matters of fact may matter, but only a little. This should be enough to get us going, but we'll come back to this issue in §2.1, where we introduce a new proposal about how similarity is determined.

Now we'll briefly outline the dynamic binding theory of indefinites and donkey pronouns we'll be using. It's based on Dynamic Predicate Logic (DPL), developed by Groenendijk and Stokhof (1991), and its extension in Groenendijk et al. (1996).[8] The approach is a dynamic one, so ultimately we'll be giving meanings in terms of an update function $[\cdot]$ which applies to an input information state from the context and returns an information state as output.

To define the notion of a state, we first need to introduce assignment functions: an assignment g is a partial function from variables to elements of the domain of individuals in the model. From this we define the notion of a possibility: a possibility i is a set of world-assignment pairs. And from this we define an information state: an information state s is a set of possibilities.

Now we'll give a partial definition of the update function $[\cdot]$ for a simple interpretation language based on FOL.[9] There are only two syntactic differences. First, the dynamic versions of the logical symbols will be marked with a ˜ above them, to distinguish them from the classical variants we are using in the metalanguage, and second, $\tilde{\exists}x$ is a well formed formula on its own. For our purposes, the important parts of the definition of the update function are the following, where $\mathcal{D}$ is the domain of the model, F is a predicate, and ϕ and ψ are formulas:

(10) a. $s[F(x)] = \{i : i \in s \wedge w_i \in [\![F(g_i(x))]\!]\}$
b. $s[\phi \tilde{\wedge} \psi] = s[\phi][\psi]$
c. $s[\tilde{\exists}x] = \{i : \exists j \exists d(j \in s \wedge d \in \mathcal{D} \wedge w_i = w_j \wedge g_i = g_j^{x \to d})\}$

An update with an atomic formula tests each input possibility for whether the formula is true at that possibility and preserves only those possibilities which pass the test. A conjunctive update is just the sequence of updates with each conjunct. And an existential update adds to the input possibilities a new possibility for each way an input possibility's assignment can be extended to provide the relevant variable with a value.

Now we will introduce an account of counterfactual donkey sentences which is a straightforward combination of the ordering semantics for counterfactuals and the dynamic binding account of indefinites and pronouns. Essentially, this is the proposal given in Wang (2009), except where she moves to a test semantics based on Veltman (2005), we'll stick more closely to the traditional idea of a counterfactual being truth conditional (which, in our dynamic framework, amounts to being eliminative).

[8] These ideas are closely related to the file-change semantics given in Heim (1982).

[9] We defined the update function directly, but we could have instead given interpretations of formulas as pairs of possibilities (input and output), then defined updates derivatively, as in DPL. Furthermore, I do not take the use of an interpretation language here to be crucial—it is used for convenience. Ultimately I'd prefer to give the semantics in a directly compositional way.

The basic idea is this: $A \boxrightarrow C$ is true at a possibility iff all the nearest A-possibilities verify C. To spell this out, we need to say what it is to be a nearest A-possibility and what it is to verify C. For any formula A, j is an A-possibility for i (or $j \in /A/_i$) iff $\exists k(g_k = g_i \wedge j \in \{k\}[A])$. So the world of j may be any world where A is true on the relevant variable assignment, but the assignment must be the result of updating i's assigment by A. A possibility is a *nearest* A-possibility (to a base possibility i) iff it is in the set that results from applying the selection function f to A and i, where

(11) $\quad f(A,i) = \{j : j \in /A/_i \wedge \neg\exists k(k \in /A/_i \wedge w_k <_{w_i} w_j)\}.$

This returns the set which includes a possibility iff it is an A-possibility (relative to i) whose world is as close to the world of i as the world of any A-possibility is. A possibility i *verifies* a formula C iff $\{i\}[C] \neq \emptyset$. That is, iff updating a state containing just that possibility with C does not lead to an empty state.

Using f to collect the nearest A-possibilities, we can give a simple dynamic ordering semantics for counterfactuals as follows:

(12) $\quad s[A \boxrightarrow C] = \{i : i \in s \wedge \forall j(j \in f(A,i) \supset \{j\}[C] \neq \emptyset)\}$

This is just the Stalnaker-Lewis idea carried over to the dynamic framework: it rules out input possibilities whose nearest A-possibilities do not all verify C.

Let's see an example of this proposal in action. We'll try it on (13), which we'll assume is the translation of our original counterfactual donkey sentence (4) into our interpretation language.

(13) $\quad (\exists x\tilde{\wedge}\texttt{donkey}(x)\tilde{\wedge}\texttt{Balaam-owns}(x)) \boxrightarrow \texttt{Balaam-beats}(x).$

Suppose we have an input context $s = \{\langle w_0, g\rangle, \langle w_1, g\rangle\}$ and a model $\mathcal{M}_1$ with the following features:

$\mathcal{I}_1$	donkey	Balaam-owns	Balaam-beats
w_0	a,b		
w_1	a,b		
w_2	a,b	a,b	a,b
w_3	a,b	a,b	a

$\mathcal{S}_1$
$w_0 <_{w_0} w_2 <_{w_0} w_3 <_{w_0} w_1$
$w_1 <_{w_1} w_3 <_{w_1} w_2 <_{w_1} w_0$

So Balaam doesn't own donkeys in either w_0 or w_1, but w_0 is closer to w_2, where he owns and beats two donkeys, than it is to w_3, where he owns two but only beats one, whereas w_1 is closer to w_3 than w_2. The semantics of (12) predicts that in $\mathcal{M}_1$, $s[(13)] = \{\langle w_0, g\rangle\}$. This is what we'd expect. An utterance of (4) should eliminate a possibility with a world like w_1, where in the nearest world where Balaam owns some donkeys, he doesn't beat all of them. And it should keep a world like w_0, where Balaam beats all the donkeys he owns in the nearest world where he owns any. So far, so good. But this theory runs into problems. Most—among them how to deal with 'weak' readings, modal subordination, and *might*-counterfactuals—I will have leave aside for now, as we turn to the one that will occupy us for the remainder of the paper.

1.1. The universal entailment problem

The problem I'd like to address is that on the semantics in (12), there is no high reading. It seems, then, not to have a way to predict the kind of entailments in (5).

Take a model, for instance, like the following:

$\mathcal{I}_2$	donkey	Balaam-owns	Balaam-beats
w_0	a,b,c		
w_1	a,b,c		
w_2	a,b,c	a,b	a,b
w_3	a,b,c	a,b,c	a

$$\mathcal{S}_2$$
$$w_0 <_{w_0} w_2 <_{w_0} w_3 <_{w_0} w_1$$
$$w_1 <_{w_1} w_3 <_{w_1} w_2 <_{w_1} w_0$$

This is like the model before, except now there's another donkey which Balaam only owns in w_3 and does not beat there. As before, in this model $\langle w_0, g\rangle$ verifies (13). But it does *not* verify (14), one of the universal entailments we would expect.

(14) $\texttt{donkey}(c)\tilde{\wedge}\texttt{Balaam-owns}(c) \,\square\!\!\rightarrow\, \texttt{Balaam-beats}(c)$

So it doesn't verify the universalized (5), either. This semantics, then, does not have high readings. How troubled should we be by this?

The first thing to note is that we don't always want these universal entailments: some counterfactual donkey sentences seem not to have them except in special contexts, and most counterfactual donkey sentences can be put in contexts where they don't seem to have these entailments. Suppose, for example, that I could pick any number between 1 and 10, but I could only pick one number. I actually picked 4. And now I say the following:

(15) If I had picked a prime number, I would have picked 2.

Certainly there is no commitment here to the truth of (16).

(16) If I had picked 7, I would have picked 2.

This was noted by van Rooij and is what WR call a 'low' reading of counterfactual donkey sentences.[10] And we might think that $\mathcal{M}_2$ and the semantics of (12) is just what we need to make sense of the following kind of argument, which includes the target counterfactual donkey sentence interpreted in a way that excludes the universal entailments.

(17) Balaam is very poor and c is a very expensive donkey. So if he owned a donkey, he wouldn't own c, but would only own a and b, who are cheap. And he would beat both a and b if he owned them. So if Balaam owned a donkey, he would beat it. That he wouldn't beat c if he owned it is irrelevant.

[10] Their accounts treat low readings in the same way as weak readings, through selective binding.

It is data like these which motivate Wang's acceptance of a semantics like (12) without high readings. But what are we to say about those cases in which the universal entailments do seem to be present? As we noted at the outset, it's pretty natural accept them when (4) is uttered without a surrounding argument like the one in (17).

1.2. The special ordering fix and WR's objection

Wang herself says nothing about how to derive these entailments. WR, however, suggest a way for an account like Wang's (and that given in (12)) to predict them in certain contexts. They observe (p. 296) that if for all the individuals in the domain, the nearest world where one individual satisfies the antecedent is no nearer than the nearest world where any other individual satisfies the antecedent, then the counterfactual donkey sentence will have the relevant universal implications. A bit more generally and formally—and this is my formulation—we get the universal entailments in contexts with a *special* similarity ordering set where

(18) A model $\mathcal{M}$'s ordering set $\mathcal{S}$ is *special* relative to an input state s iff
$\forall i(i \in s \supset \forall j(j \in /A/_i \supset \exists k(k \in f(A,i) \wedge g_j = g_k)))$.

That is, for all possibilities i in s, if j is an A-possibility for i, then among the nearest (relative to i) A-possibilities is a possibility which shares an assignment with j.

For example, in a model with the interpretation $\mathcal{I}_2$ and input state as before, what would the ordering set have to look like in order for it to be special? The ordering for w_0 in $\mathcal{S}_3$ is $w_0 <_{w_0} w_2 <_{w_0} w_3 <_{w_0} w_1$. This prevents $\mathcal{S}_2$ from being special, since $\langle w_0, g\rangle$ is in s and there is a possibility in $/A/_{\langle w_0,g\rangle}$, namely $\langle w_3, g^{x\to c}\rangle$, that does not share an assignment with any possibility in $f(A,\langle w_0,g\rangle)$, which only includes $\langle w_2, g^{x\to a}\rangle$ and $\langle w_2, g^{x\to b}\rangle$. To make the ordering special, we need to adjust the ordering of worlds so that $f(A,\langle w_0,g\rangle)$ also includes a possibility whose assignment is $g^{x\to c}$. Since w_3 is the only world where c is a donkey owned by Balaam, we can only do this by making $w_3 \leq_{w_0} w_2$. For illustration, let's let the new ordering for $\mathcal{S}_4$ be $w_0 <_{w_0} w_2 =_{w_0} w_3 <_{w_0} w_1$. We can keep $<_{w_1}$ as before.

Now that we have a special ordering, let's see how it generates the universal entailments. The trouble we ran into before is that $\langle w_0, g\rangle$ verified (13) but not (14), repeated here as (19a) and (19b), respectively.

(19) a. $(\exists x \tilde{\wedge} \texttt{donkey}(x) \tilde{\wedge} \texttt{Balaam-owns}(x)) \,\square\!\!\rightarrow\, \texttt{Balaam-beats}(x)$.
b. $\texttt{donkey}(c) \tilde{\wedge} \texttt{Balaam-owns}(c) \,\square\!\!\rightarrow\, \texttt{Balaam-beats}(c)$

But now with $f(A,\langle w_0,g\rangle)$ including $\langle w_3, g^{x\to c}\rangle$, it will no longer be that $\langle w_0, g\rangle$ verifies (19a), since $\{\langle w_3, g^{x\to c}\rangle\}$ doesn't verify C. And to get an interpretation on which it *does* verify (19a), Balaam would have to beat c in w_3.

To keep the semantics as is, we need to have a special ordering to get the universal entailments. What the simple ordering semantics + dynamic binding theory predicts, then, is that the uni-

versal entailments only arise in contexts where the similarity ordering is special. WR, however, claim to empirically falsify this prediction, and reject Wang's account on these grounds. They claim that there are contexts on which the entailments arise but on which the similarity ordering is not special. I will illustrate their point with my own case, but it's in the same spirit as the one they offer.

(20) SCENARIO: Balaam took part in a game show which had the following format: if you win the easy first round, you win Herbert, an obnoxious and disobedient donkey. The reward for the much more difficult second and third rounds are the well-mannered and obedient donkeys Eeyore and Platero, respectively. Losing a round of the game eliminates the player, keeping them from advancing to any later rounds. Balaam was eliminated in the first round, and so remains donkeyless.

John, only aware of the game's first round, asserts our original counterfactual donkey sentence (4), repeated here as (21), since he knows about Balaam's short temper.

(21) If Balaam owned a donkey, he would beat it.

Sarah, who has more information about the game, corrects him with (22).

(22) No, Balaam could have won Platero or Eeyore too, and he wouldn't beat either of them if he owned them.

It is implausible, WR would contend, to claim that in this context a world where Balaam advances to and wins the third round is just as similar to the actual world as the one where he wins just the first round. But this is what we would have to say to give John's utterance a false reading which (22) can be used to disagree with.

$\mathcal{I}_3$	`donkey`	`Balaam-owns`	`Balaam-beats`
w_0	h, e, p		
w_1	h, e, p	h	h
w_2	h, e, p	h, e	h
w_3	h, e, p	h, e, p	h

$\mathcal{S}_3$-Intuitive

$w_0 <_{w_0} w_1 <_{w_0} w_2 <_{w_0} w_3$

$\mathcal{S}_3$-Special

$w_0 <_{w_1} w_1 =_{w_1} w_2 =_{w_1} w_3$

To get the needed universal entailments, we need $\mathcal{S}_3$-Special, but it's hard to see how the ordering could be like that, rather than like the non-special $\mathcal{S}_3$-Intuitive.

1.3. A semantics with high readings

WR conclude that van Rooij was right: we need to give a semantics of counterfactual donkey sentences which has a reading—the high reading—on which the universal entailments arise no matter what the similarity ordering, no special order needed.

The main innovation in van Rooij's proposal is to derive similarity orderings over possibilities out of the ones over worlds, and allow possibilities to be comparable only if they share an assignment function. We define an assignment-sensitive similarity ordering $\leq^*$ based on the old world ordering $\leq$ as follows.[11]

(23) $j \leq^*_i k$ iff $w_j \leq_{w_i} w_k \wedge g_j = g_k$

Now the only changes we need to the semantics is to have a selection function use this assignment-sensitive ordering rather than the ordering on worlds, and a semantics for $\Box\!\!\rightarrow$ which uses the new selection function.

(24) $f^*(A,i) = \{j : j \in /A/_i \wedge \neg\exists k(k \in /A/_i \wedge k <^*_{w_i} j)\}$
(25) $s[A \Box\!\!\rightarrow C] = \{i : i \in s \wedge \forall j(j \in f^*(A,i) \supset \{j\}[C] \neq \emptyset)\}$

What this new assignment-sensitive semantics does is encode the need to check for each assignment the nearest A-possibility with that assignment whether it verifies C, regardless of whether there are A-possibilities with different assignments and nearer worlds. This predicts the universal entailments regardless of the similarity ordering on worlds. And in particular, it predicts the universal entailments for (21) in scenario (20) without having to posit the supposedly implausible special ordering. Using $\mathcal{S}_3$-Intuitive, the assignment-sensitive selection function f^* will return $\{\langle w_1, g^{x\to h}\rangle, \langle w_2, g^{x\to e}\rangle, \langle w_3, g^{x\to p}\rangle\}$, which has members (namely $\langle w_2, g^{x\to e}\rangle$ and $\langle w_3, g^{x\to p}\rangle$) which do not verify $\texttt{Balaam-beats}(x)$. So the assignment-sensitive semantics predicts the sentence to be false in this model, as desired. And more generally, a possibility will verify a counterfactual donkey sentence iff it also verifies the universal entailments. Thus WR claim that this kind of case supports assignment-sensitive semantics like (25) over the assignment-insensitive theories like the one given in (12).[12]

2. Why we don't need the high reading

I find WR's argument for the assignment-sensitive semantics unpersuasive. As mentioned in §1, we can't always assume that the similarity ordering in the semantics of counterfactuals matches an intuitive notion of similarity. So we can't just appeal to our intuitive idea of similarity, as WR do, to rule out the special ordering in scenario (20). Later in this section I will sketch what we need to say about the similarity relation to get the special ordering in the relevant scenarios, and defend this as a tenable view. But first I will argue that even if we move to the assignment-sensitive semantics with high readings, we still need to appeal to an ordering of the special kind in scenarios just like (20) in order to correctly predict the presence of *strong* entailments.

The takeaway is that within the ordering semantics + dynamic binding framework, the move to assignment-sensitivity doesn't save us from having to appeal to a special ordering in scenarios

[11] In van Rooij's original formulation and WR's follow-up, there's an additional condition on the similarity orderings for possibilities: for $\leq^*_i$ to hold between j and k, it must be that $g_j = g_k \supseteq g_i$. As far as I can tell, though, this doesn't play any helpful role.

[12] It is only after they make this argument that they move, for independent reasons (based on NPI data), to a dynamic strict theory. As they present the argument discussed in this section, it is an argument in favor of van Rooij's account (more or less that of (25)) over Wang's.

like (20) anyways, so, at least in debates between those who share this framework, the proponent of assignment-*in*sensitive semantics may avail herself of this ordering to generate the universal entailments we've been discussing.[13]

To make this point, we need to make a distinction between two kinds of universal entailments a counterfactual donkey sentence might have. The universal entailments we've been discussing so far, the *high* entailments, are like those we'd expect from universal quantification scoping over the whole conditional—everything is such that if it were a donkey Balaam owned, he would beat it. This is as opposed to *low* entailments, which can be true so long as the conditional is true of the thing that would satisfy the antecedent, were it to hold. The other kind of universal entailments are *strong* entailments, which are like those we'd expect from universal quantification in the consequent—if Balaam owned a donkey, he would beat every donkey he owned. This is as opposed to *weak* entailments, which do not require this. We can summarize these entailments with (partial) paraphrases:

High	Low
$\forall x((A(x) \boxrightarrow C(x))$	$(\exists x A(x)) \boxrightarrow C(x))$

Strong	Weak
$\ldots \boxrightarrow \forall x(A(x) \supset C(x))$	$\ldots \boxrightarrow \exists x(A(x) \wedge C(x))$

Once we've made these distinctions, we can see first, that strong and weak can each be combined with high and low, and second, that these combinations are not equivalent with the simple high and low entailments as stated above.

	High	Low
Strong	$\forall x((A(x) \boxrightarrow \forall y(A(y) \supset C(y)))$	$(\exists x A(x)) \boxrightarrow \forall y(A(y) \supset C(y)))$
Weak	$\forall x((A(x) \boxrightarrow \exists y(A(y) \wedge C(y)))$	$(\exists x A(x)) \boxrightarrow \exists y(A(y) \wedge C(y)))$

What is important for us are the contrasts between high/weak, high/strong, and simple high entailments. The contrast between high/weak and the others is easy enough to see; note that in the problem case from §1.2, $\mathcal{M}_3$ with $\mathcal{S}_3$-Intuitive, the high/weak entailments for (4) in fact hold, though the high and high/strong ones don't. The high/strong vs. high contrast is less obvious, but in the next section we'll see an example that makes it clear that they can differ.

The problem for the assignment-insensitive semantics was supposed to be that it failed to predict high entailments given an intuitive similarity ordering, and the proposed solutions were special orderings on the one hand and moving to assignment-sensitivity on the other. But while assignment-sensitivity does, on its own, get us simple high entailments, it doesn't get us high/strong entailments. It turns out that to get high/strong entailments, the assignment-sensitive semantics also needs to use special orderings. But the special orderings can get us the high/strong entailments without assignment-sensitivity. So if we want high/strong entailments,

[13] Actually, this is a bit stronger conclusion than is warranted. Perhaps there are other semantic theories which still make use of this framework, but work out the details in a different way. And perhaps there could be some reasonably non-ad hoc such theory that is assignment-sensitive and gets both high and strong readings.

rather than just high entailments, it seems that we're going to need to appeal to special orderings. Let's illustrate this point with a case for which we need high/strong, and not merely high, entailments.

(26) SCENARIO: Cory, who is donkeyless, is a bit crazy. He's disposed to take out his anger on his most prized possession. He also took part in the game show described in (20), but also lost in the first round. Had he won any rounds, the prize from the most advanced round he won would have become his prized possession, and he would have beaten it, but he wouldn't beat anything else.

Now consider the following:

(27) If Cory owned a donkey, he would beat it.

In this scenario, the salient reading of (27) seems false. And it seems false because the relevant high/strong entailments don't hold. If Cory had owned Eeyore, for example, it wouldn't be true that he would beat every donkey he owned. But note that all of the high entailments *do* hold. If Cory owned Herbert, he would beat Herbert; if he owned Eeyore, he would beat Eeyore; and if he owned Platero, he would beat Platero. So, first point from this example: high/strong $\neq$ high.

The structure of this scenario is very similar to that of scenario (20). Again we can consider models with the intuitive ordering and a corresponding special ordering

$\mathcal{I}_4$	`donkey`	`Cory-owns`	`Cory-beats`
w_0	h, e, p		
w_1	h, e, p	h	h
w_2	h, e, p	h, e	e
w_3	h, e, p	h, e, p	p

$\mathcal{S}_4$-Intuitive
$w_0 <_{w_0} w_1 <_{w_0} w_2 <_{w_0} w_3$

$\mathcal{S}_4$-Special
$w_0 <_{w_1} w_1 =_{w_1} w_2 =_{w_1} w_3$

Second point from this example: to get the desired high/strong entailments, both the assignment-insensitive semantics in (12) and the revised, assignment-sensitive semantics from (25) need to use a special ordering, such as $\mathcal{S}_4$-Special. Using $\mathcal{S}_4$-Intuitive, the assignment insensitive f will return $\{\langle w_1, g^{x\to h}\rangle\}$, and the assignment-sensitive f^* will return $\{\langle w_1, g^{x\to h}\rangle, \langle w_2, g^{x\to e}\rangle, \langle w_3, g^{x\to p}\rangle\}$. In either case, all of the selected possibilities verify `Cory-beats`(x), yielding the prediction that the counterfactual is true on either semantics. So no high/strong entailments either way.

But when we move to $\mathcal{S}_4$-Special, both selection functions will return $\{\langle w, h\rangle : w \in \{w_1, w_2, w_3\} \wedge h \in \{g^{x\to h}, g^{x\to e}, g^{x\to p}\}\}$, predicting a *false* reading given $\mathcal{I}_4$, and more generally a false reading unless the strong/high entailments also hold.

Just as with the case in §1.2, the special ordering is what we need to get the right prediction for the assignment-insensitive semantics, despite the special ordering not matching the intuitive one for the scenario. But now this is also what we need to make the right prediction for the assignment-sensitive semantics as well. The move to assignment-sensitivity, then, doesn't keep

us from needing to appeal to special orderings in these sorts of scenarios. This undermines WR's argument for assignment-sensitivity on the grounds that it does avoid such appeals.

2.1. Why the special ordering?

Perhaps, though, what we should conclude from this not that WR's argument doesn't refute the assignment-insensitive semantics, but that it *also* refutes the assignment-sensitive semantics, and that a more radical revision is required.[14] We should require a high/strong reading, one that predicts these entailments even without a special ordering.

In this section I want to briefly defend the view that we need not make such a move—that using the special ordering even in the scenario is not particularly implausible. I do so by outlining a proposal about the similarity relation which will predict special orderings in the relevant contexts. This will not yet give us an argument against accounts which get the entailments through a high/strong reading instead of through the special ordering; that argument will come in §3.

What makes one world closer than another to an evaluation world? As discussed in §1, the answer can't simply be that considering all the facts in these worlds, it is more intuitively similar than the other is. For the same reason, it can't be that it has some greater amount (by some measure) of overlap in facts, where all facts count the same. Facts of some kinds count more heavily in determining (dis)similarity than others.

However, proposals like Lewis's (as well as others in the same spirit) to weigh differences in certain kinds of fact more heavily than others won't predict the special ordering in all the cases we would need special orderings for. We may assume that it takes more widespread miracles and less perfect match of particular facts with the evaluation world for Balaam or Cory to win two or three rounds of the game than it would for them to win just one. We should say, then, what an account of the similarity relation would have to look like to get the special orderings when we need them.

Here is my suggested amendment: we start with some standard account, such as Lewis's, for determining similarity. Then we allow the similarity orderings to be affected by the antecedent of an asserted counterfactual. In particular, we say that for each of the salient ways the antecedent might be made true, how it is made true in a given world is irrelevant to determining similarity.[15] When determining similarity between two worlds, we look for violations of law, amount of mismatch in particular fact, and so on, *except in the parts of the world that are involved in making the antecedent true*.

[14] I suspect WR would be sympathetic to this extension of their argument, since their independently motivated move to a dynamic strict account avoids the problem with high/strong entailments. Their account predicts these, rather than the merely high entailments that they claim it does.

[15] Regarding 'making true', I have in mind something along the lines of Fine's notion of exact truthmaking (see Fine (2017)). 'Salient' is, as it often is, left vague and underdeveloped. Investigating what account of salience gets us the best results for this proposal would be a worthwhile undertaking, but not one I can pursue here.

This will typically result in special orderings for counterfactuals with indefinites in the antecedent. If the reason that the world where Cory owns Eeyore and Platero as well as Herbert was farther from the evaluation world than the one where he owns just Herbert is that it takes more widespread miracles for him to own all three than it does to just own Herbert, then these worlds will be brought to the same degree of similarity by my proposal, since it tells us to ignore the differences involved in the salient ways of making it true that Cory owns a donkey, and his owning the three donkeys is one salient way of owning a donkey and his owning just Herbert is another.

We need not take this proposal to be entirely ad hoc, since it extends beyond counterfactuals with indefinites in the antecedent to any counterfactual with an antecedent that has more than one salient alternative way which might make it true. This gives us nice results both for counterfactuals with disjunctive antecedents as well as other unspecific antecedents.

Since asserting counterfactuals with disjunctive antecedents presumably makes salient the possibility of either disjunct making the antecedent true, our proposal tells us that which disjunct is true (and what goes into making it true) is irrelevant to similarity, regardless of differences in amount of miracles required to make each true. Thus, we'd expect the conjunctive implications from counterfactuals with disjunctive antecedents in the cases that motivate acceptance of simplification of disjunctive antecedents. For instance, we would expect to be able to infer (28b) from (28a) in the example from Nute (1975), since according to my proposal there should be worlds where the sun grows cold that are just as similar to the actual world as any world where we have good weather.

(28) a. If we were to have good weather this summer or if the sun were to grow cold before the end of the summer, we would have a bumper crop.
b. If the sun were to grow cold before the end of the summer, we would have a bumper crop.

For a case without disjunction or an indefinite, but with multiple salient ways of making the antecedent true, consider this example from Bennett (2003: 219–220), who attributes the idea to John Pollock as reported by Nute (1980: 104).

(29) SCENARIO: My coat was not stolen from the restaurant where I left it. There were two chances for theft—two times when relevant indeterminacies or small miracles could have done the trick. They would have involved different potential thieves; and the candidate for the later theft is a rogue who always sells his stuff to a pawnbroker named Fence.
(30) If my coat had been stolen from the restaurant, it would now be in Fence's shop.

Our first reaction is that in this scenario this counterfactual is false, or at least unassertable, since it might have been stolen by the earlier thief. But a straightforward application of a proposal like Lewis's predicts that it would be true, since the later theft would involve a larger region of perfect match of particular fact. With my proposal, though, we ignore this difference in fact-matching, since they're involved in salient ways of making the antecedent true. This

means that there will be among the closet worlds worlds where the earlier thief steals my coat and worlds where the later thief steals it. We thus predict (30) to be false. We get a similar desirable result if we modify the example to let one of the salient possible thefts involve a bigger miracle, rather than just an earlier one. And having seen this pattern, examples can be easily multiplied.

These seem like attractive results. However, the proposal won't work for all cases. As is well known from work on disjunctive antecedents, and as we noted in §1, there are exceptions to these patterns. In some contexts, apparently it *does* matter how the antecedent is made true.

(31) a. If Spain had joined the Allies or the Axis, they would have joined the Axis.
b. If Spain had joined a side, it would have been the Axis.

I assume that in these cases the alternative of joining the Allied side will be salient in whatever sense of salience we need for the special ordering. But for them to be true, as they plausibly are, we can't have worlds where Spain joins the Allies rather than the Axis as among the closest worlds where the antecedents are true.

In addition, even for sentences which have high entailments by default, we can construct contexts in which these entailments do not arise.

(32) If Balaam had won any rounds, he would have won just the first one. So if he had owned a donkey, he would have only owned Herbert. And if he had owned Herbert, he would have beat him. So if he owned a donkey, he would beat it.

In this context, the sentence seems true, even though the high entailments still fail to hold. So not only do we need an account that is sometimes indifferent to antecedent truthmaker, we also need one that is sometimes *not* indifferent.[16] How similarity is determined, then, depends on whether the utterance is in an antecedent-truthmaker-relevant or -irrelevant context. Ultimately we'll want an account of what makes a context one way or the other, and what unified account of similarity, if any, underlies them. But for now we'll just accept that there are these two types of orderings, and one type—the antecedent-truthmaker-irrelevant—typically leads to special orderings for counterfactual donkey sentences.[17]

[16] Alternatively, we might try to capture these differences as differences in 'salience' of alternatives, a suggestion made to me by Kyle Blumberg. As much of a black box as it currently is, we might be able to construe 'salience' in a way that allows for this. However, it won't be easy, since to get the identificatory disjunctive antecedent cases right, we'd have to say that these antecedents don't make both disjuncts salient.

[17] This bifurcated account is somewhat like the orderings that would be required to make sense of backtracking as well as non-backtracking counterfactuals.

(i) a. If he jumped, he would have died.
b. He wouldn't have jumped unless there was a net. So if he had jumped there'd be a net there to save him and he wouldn't have died.

Indeed, it's not so far-fetched to think that non-backtracking and antecedent alternative indifference on the one hand, and backtracking and antecedent alternative sensitivity on the other are instances of the same phenomenon. Perhaps backtracking counterfactuals are those which the truthmaker of the antecedent is relevant to similarity, on a construal of truthmaker which includes causal origins.

This is only a sketch of an account, but it seems to me not an obvious dead-end. For WR's argument to succeed, we would need to show that nothing along these lines of determining similarity could work, since we would need to rule out special orderings in the relevant cases. So until that's done, we should take the assignment-insensitive semantics with special orderings to be a viable account of counterfactual donkey sentences with high entailments.

Before proceeding to the final section, I wish to point out one consequence of the current special ordering-based account. If, as I think we should, we still require the similarity relation to be strongly centered, even in the antecedent-truthmaker-irrelevant contexts, we will not always be able to produce a special ordering, and so won't be able to generate high entailments. This is because in some cases, the antecedent will be true *in the evaluation world*. Since the evaluation world will be strictly closer to itself than any other world is, worlds with different ways of making the antecedent true will not be as close, preventing a special ordering. In these cases, on the assignment-insensitive ordering semantics, we shouldn't expect high entailments.[18] This prediction differs from that of accounts which would allow for high (or high/strong) readings. On such views, we should expect there to be high entailments in some of these cases, since the high entailments are not dependent in any way on the similarity ordering of worlds.

I will argue that on this point, the assignment-insensitive ordering semantics make the correct predictions and accounts with high (or high/strong) readings make incorrect ones. There are no high entailments when the antecedent is true.

3. Why we don't want the high reading

Before looking at a case where the predictions of the different accounts come apart, I'd like to make an observation about what is involved in high entailments. We've been putting them in terms of a universal quantifier, ranging over a domain given in the model, implicitly assumed to be restricted by the context.[19] But we've glossed over what exactly this amounts to by only looking at cases where what exists does not vary from world to world.[20] Once we start to look at scenarios where this assumption is dropped, we need to ask whether the outermost $\forall$ used to state what the high entailments are is meant to range over just those things that exist in the world of evaluation, or rather whether it includes merely possible entities as well. In other words, we need to ask whether this is an actualist or possibilist quantifier.

I think it's clear enough how this question is to be answered. Suppose Allie and Bert think Mary the potter probably didn't make anything yesterday. And now Allie says the following:

[18] Why don't we give up strong centering for these cases? After all, if the only difference between the evaluation world and some other one is how the antecedent is made true, we might expect them to be equally similar to the evaluation world. This is worth considering, but the data in the next section suggest we should not do so.

[19] Most likely it's a bad idea to put this contextual restriction in the models themselves, rather than, say, putting domain variables in the syntax. See Stanley and Szabó (2000). Incidentally, we might want our similarity orderings not to be just given in the models either, but at least partly determined through something syntactically present. One proposal for how to do this is made in Arregui (2009).

[20] And in particular where the extension of the restrictor of the antecedent's indefinite (in the cases we've looked at: what is a donkey) doesn't vary from world to world.

(33) If Mary had made a vase, she would have made it from glass.

Now consider:

Case 1: Mary didn't make any vases, and there is no contextually relevant actual pottery.

In this case, (33) does not come out trivially true or give rise to presupposition failure. Nor does it depend on looking at various non-pottery that exists in the world (the high entailment need not imply, for example, that if Mary were a vase that she made, she would have made herself from glass). Instead, what goes into determining the truth or falsity of (33) in Case 1 are some merely possible vases and their composition in worlds where Mary made them—the quantifier in question is a possibilist one. So if there are high readings of counterfactual donkey sentences, they require that all relevant possible entities are such that, in the closest worlds where they satisfy the antecedent, they satisfy the consequent.

With this in mind, let's return to the different predictions made by the special ordering and the high reading accounts of high entailments. If there are high readings, we should expect high entailments even when the antecedent is true in the evaluation world, and we've now seen that this requirement extends to merely possible entities. If high entailments come from a special similarity ordering, we shouldn't expect them to arise when the antecedent is true in the evaluation world, since in such cases there can be no special ordering without violation of strong centering.

Let's look a case, then, where some contextually relevant actual entity satisfies the antecedent and consequent but a merely possible one doesn't satisfy the consequent in the nearest world where it satisfies the antecedent. Suppose the conversation between Allie and Bert continues:

(34) a. *Bert*: No, she could have made it from clay!
b. *Allie*: Oh, I didn't know she had any clay left, nevermind what I just said, then.

So here Bert raises a relevant possible way for the antecedent to be made true that wouldn't lead to the truth of the consequent, which gets Allie to retract her claim. This is just what we'd expect on a high reading. But now suppose that it turns out that Mary in fact *did* make some vases yesterday.[21]

Case 2: Mary made two vases, both of glass.

In this case it seems that Allie's utterance of (33) was true, if only by luck. And the fact that Mary could have made a different vase and she would not have made that one from glass has no

[21] One might worry that this would make the original assertion infelicitous or at least difficult to evaluate, since *counter*factuals generally presuppose the falsehood of their antecedent. But there are well known exceptions to this generalization, such as the famous case from Anderson (1951): "If Jones had taken arsenic, he would have shown just exactly those symptoms which he does in fact show." That Allie and Bert are unaware at the time of utterance that the antecedent is true, but are also not certain that it's false, should make it clear enough that nothing too strange is going on here.

bearing on its truth. Once the facts about what Mary actually made are known to Allie and Bert, challenging the original assertion again by raising the possibility of the clay vase is bizarre.

(35) a. *Allie*: Looks like I was right after all.
b. *Bert*: ??No, even though she *didn't* make any clay vase, she still *could* have made a vase from clay, and she wouldn't have made *that* from glass.

So it seems that in this case, any available reading of (33) is true, regardless of the possibility Bert raises. The high entailments seem not to arise in this case. This is just as the special ordering (with strong centering) account predicts. And it's not predicted by accounts which allow high readings. Where there are high entailments, I conclude, we should take them to be due to a special ordering rather than being baked in semantically, through a high reading.

There are a couple objections we should address. First, the proponents of high readings—van Rooij and WR—allow that in some contexts there are low (or for them, low/weak) readings. Why not think this is what's happening here? Two reasons. First, because their method for obtaining low readings guarantees weak readings. But the salient reading of (33) is not weak. Consider:

Case 3: Mary made one vase of glass and one of clay.

Here we would take (33) to be false, even though it would be true in Case 2. But on a weak reading, it would also be true in Case 3.

Second, this utterance seems like it has high entailments in the evaluation worlds where the antecedent is not true (like in Case 1). This is why Allie retracts her assertion once the possible clay vase is brought to her attention—for all she knows, she and Bert are in a world where Mary made no vases, and in such a world her assertion, if it had high entailments, would be false. It's difficult to see how an account with high readings would treat this utterance as having a high reading in Case 1, but a low reading in Case 2.

The other objection is that the merely possible clay vase gets ignored due to quantifier domain restriction. To evaluate this properly, it would be important to spell out what the account of domain restriction would have to look like to get this right. But there is some reason ahead of time to doubt that it would work, given our judgments of the other cases. A possible clay vase needs to be in the domain to make Bert's original interjection true, and it seems that it is deemed relevant by both Bert and Allie to Allie's claim. So we would need to have an account which does not exclude this possible vase through quantifier domain restriction when it should be included—Case 1, for example—and exclude it when there happens to be actual vases that Mary made. I don't know how this could be done in a way that's not implausibly ad hoc.

It could be that either of these or some other objection could be worked out together with an account with high (or better, high/strong) readings that treats the above cases successfully. But given the difficulties that would seem to involve, and the fact that the data in this section is just what the special ordering account predicts, I tentatively conclude that counterfactual

donkey sentences do not get high readings, but instead get their occasional high entailments from special orderings.

References

Anderson, A. R. (1951). A note on subjunctive and counterfactual conditionals. *Analysis 12*(2), 35–38.

Arregui, A. (2009). On similarity in counterfactuals. *Linguistics and Philosophy 32*, 245–278.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

Carter, S. and S. Goldstein (ms). Might counterfactual donkey sentences.

Fine, K. (1975). Critical notice of David Lewis's *Counterfactuals*. *Mind 84*, 451–458.

Fine, K. (2017). Truthmaker semantics. In B. Hale, C. Wright, and A. Miller (Eds.), *A Companion to the Philosophy of Language* (2 ed.), Volume 2. Wiley Blacwell.

Groenendijk, J. and M. Stokhof (1991). Dynamic predicate logic. *Linguistics and Philosophy 14*, 39–100.

Groenendijk, J., M. Stokhof, and F. Veltman (1996). Coreference and modality. In S. Lappin (Ed.), *The Handbook of Contemporary Semantic Theory*, pp. 179–213. Blackwell.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. Ph. D. thesis, University of Massachusetts, Amherst.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof (Eds.), *Formal Methods in the Study of Language*, pp. 277–322.

Lewis, D. (1973). *Counterfactuals*. Blackwell.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs 13*, 455–476.

Nute, D. (1975). Counterfactuals and the similarity of words [*sic*]. *The Journal of Philosophy 72*(21), 773–778.

Nute, D. (1980). *Topics in Conditional Logic*. D. Reidel Publishing Company.

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in Logical Theory*. Oxford University Press.

Stanley, J. and Z. G. Szabó (2000). On quantifier domain restriction. *Mind & Language 15*(2), 219–261.

van Rooij, R. (2006). Free choice counterfactual donkeys. *Journal of Semantics 23*(4), 383–402.

Veltman, F. (2005). Making counterfactual assumptions. *Journal of Semantics 22*, 159–180.

von Fintel, K. (2001). Counterfactuals in a dynamic context. *Current Studies in Linguistics 36*, 123–152.

Walker, A. (2017). *The world is not enough: situations, laws and assignments in counterfactual donkey sentences*. Ph. D. thesis, Universität Konstanz.

Walker, A. and M. Romero (2015). Counterfactual donkey sentences: A strict conditional analysis. *Proceedings of SALT 25*, 288–307.

Wang, Y. (2009). Counterfactual donkey sentences: a response to Robert van Rooij. *Journal of Semantics 26*(3), 317–328.