Subjectivity in *Force*: Topicalization, Context-Sensitivity, and Morality

Matthew Mandelkern and Jonathan Phillips

1. Overview: Research has shown that moral judgments have a surprising impact on an array of distinct phenomena, e.g. judgments of causation and intentional action (e.g. Knobe (2010)). The impact of morality on these phenomena has been shown to be subject to order-effects and exhibit substantial variation between individuals (e.g. Feltz and Cokely (2011)), suggesting that such effects arise due to the impact of an underlying subjective dimension. We propose to get a better understanding of these phenomena by making a case study of the impact of morality on judgments about *force*.

2. Baseline: To do so, we conducted a series of experiments ($N = 1589$). We first used a scenario outlined in (1) to replicate previous research showing participants agree at lower rates that someone was forced to do φ when φ is morally bad than when φ is neutral or good (Phillips and Knobe, 2009; Young and Phillips, 2011).

(1) A captain overtaken by a storm realized his ship would sink if he didn't make it lighter. The only things on the boat were a sailor, small but expensive cargo, and several passengers. The captain ordered the sailor to throw the [cargo/passengers] overboard. The [cargo/passengers] sank to the bottom of the sea. The captain and his ship survived.

We replicated the previously observed pattern: Participants in the morally neutral (cargo) condition agreed with (2) at higher rates ($M = 5.37$) than those in the morally bad (passenger) condition ($M = 3.54$, $t(195.53) = 8.55$, $p < .001$, $d = 0.87$ (on a 1-7 Likert agreement-scale)).

(2) The sailor was forced to throw the [cargo/passengers] overboard by the captain.

2. Topicalization: We then tested the effect of topicalizing either the forcer or forcee, while holding fixed the action. We tested this with passivization, as in (3-a)/(3-b), and with 'as for' topicalization.

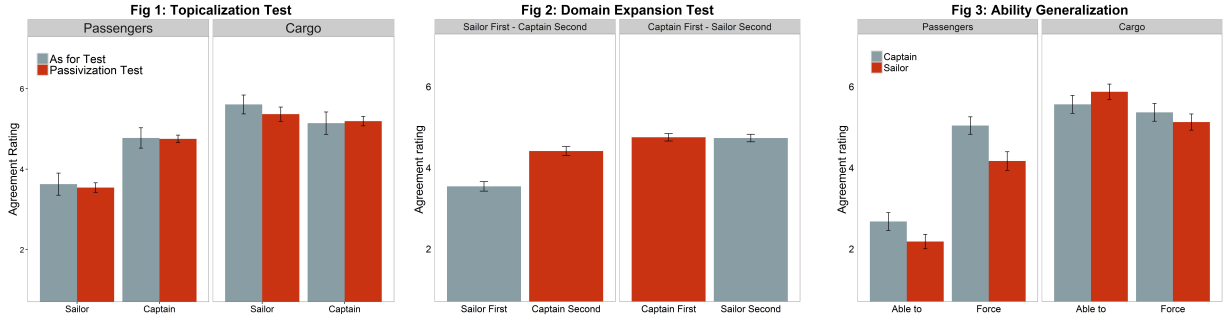
(3) a. The captain forced the sailor to throw the [cargo/passengers] overboard.
b. The sailor was forced to throw the [cargo/passengers] overboard by the captain.

We observed an interaction effect between topicalization and the morality of the action ($F(1, 1278) = 37.95$, $p < .001$, $\eta_p^2 = .027$): When the action was immoral, participants more agreed when the captain was topicalized than when the sailor was ($t(779.22) = 8.63$, $p < .001$, $d = 0.602$). This pattern disappeared in the non-moral variant ($t(443) = -1.46$, $p = .146$, $d = 0.146$; see Fig. 1).

Topicalization is known to affect how context determines quantifier domains (von Fintel, 1994). As a preliminary hypothesis, we thus propose to explain this effect by positing that the meaning of 'force' varies with a contextually given domain, whose assignment is affected by topicalization.

3. Order effects: Quantifier domain expansion is 'sticky': when a domain is expanded, it tends to remain large, rather than reverting to its original size (Lewis, 1979; von Fintel, 2001). A quantifier domain hypothesis thus yields a prediction: if A and A' mean the same thing, except that A' expands a quantifier domain, then, in a sequence of assertions $\langle A, A' \rangle$, rates of agreement with the first assertion will differ from rates of agreement with the second. By contrast, in a sequence $\langle A', A \rangle$, rates of agreement will stay constant, since the first assertion introduces an expanded domain which persists for the second. Thus, to test our hypothesis about 'force', we tested rates of agreement with sequences of topicalized force sentences. The first sequence ran $\langle$(3-b), (3-a)$\rangle$, topicalizing the

sailor, then captain. The reverse sequence, $\langle$(3-a), (3-b)$\rangle$, topicalized the captain, then sailor. Our results showed a striking pattern (see Fig. 2). In the first sequence, subjects have low agreement with (3-b), then higher agreement with (3-a). By contrast, in the second sequence, subjects have high and uniform agreement for both sentences. Statistically, this is captured by a significant interaction effect between the topic of the sentence and the order of presentation, $\chi^2(1) = 54.417$, $p < .001$.



4. Analysis: These results confirm the pattern expected from quantifier domain expansion, and provide two further data points. First, since higher rates of agreement are the 'stickier' ones, domain expansion in this case results in *elevated* rates of agreement. Second, since the captain-topicalized sentence induces 'sticky' rates of agreement, it is topicalizing the *captain* which forces domain expansion. We propose to make sense of this as follows. First, we adopt this semantics for 'force':

(4) $[\![$A forced B to $\varphi]\!]^{c,w}$=1 iff the following are true in $\langle c, w\rangle$: (i) 'B did $[\![\varphi]\!]^c$'; (ii) 'B was unable to refrain from $[\![\varphi]\!]^c$'; (iii) 'A made (ii) true'; and (iv) 'If (ii) had not been true, then B would not have done $[\![\varphi]\!]^c$'.

This semantics, as desired, incorporates a context-sensitive quantifier, in the ability modal in (ii). But if we adopt the standard treatment of 'able' as an existential quantifier over worlds (Kratzer, 1977, 1981), we face a puzzle: the negated ability claim in (ii) will have *universal* force, and so expanding the relevant domain should lead to *depressed* levels of agreement, contrary to observation.

We avoid this puzzle by adopting a non-standard semantics for 'able'. We modify the approach in (Mandelkern et al., 2015, 2016), who outline independent problems for the standard semantics and propose an alternative: With $f_c(w, \psi)$ (Stalnaker, 1968)'s selection function, taking a world and proposition to the closest world where that proposition is true; with $\mathcal{C}_{c,w}$ a *causal background*, which we model as a set of propositions; and with $\Pi_c(\cap\mathcal{C}_{c,w})$ a contextually salient cover of $\cap\mathcal{C}_{c,w}$, which we interpret as the set of salient actions compatible with the relevant causal background:

(5) $[\![$S is able to $\varphi]\!]^{c,w}$=1 iff $\exists A \in \Pi_c(\cap\mathcal{C}_{c,w}) : [\![$S does $\varphi]\!]^{c,f_c(w,[\![S\ tries\ to\ do\ A]\!]^c)} = 1$.

Informally: $\ulcorner$S is able to $\varphi\urcorner$ is true just in case, for some action A, made salient by context and compatible with the causal background, S does φ in the closest world where S tries to do A.

What is crucial about this semantics is that, unlike the standard one, it predicts that *expanding the relevant domain* – here, $\mathcal{C}_{c,w}$ – will make the set of relevant actions *smaller*, making it *harder* for ability ascriptions to be true as the domain expands, and thus *easier* for force ascriptions to be true. This is consistent with our results, which, again, suggest that force ascriptions become more plausible when the relevant domain grows. And it affords a substantive explanation of our results: We propose that, when we are thinking about moral agency, we tend to ignore substantial

parts of a causal background, conceptualizing agents as free to act in morally *good* ways. Focusing our attention on an agent by topicalizing them, as in (3-b), heightens this tendency. By contrast, however, when we make salient elements of the causal background − by topicalizing the captain, reminding the interlocutors of his causal role − it becomes harder to ignore those causal factors. The set of background causes thus expands, making 'force' judgments more likely to be true.

An alternate hypothesis would be that morality's influence is exerted through the third component of our analysis − that A made (ii) true − which is itself a causal matter. This would not on its own explain the quantifier domain pattern, though perhaps a story could be worked out. To test this hypothesis, we conducted a final experiment, which asked subjects directly about ability instead: given scenario (1), we asked whether the sailor was able to do anything other than throwing the passengers/cargo overboard. We found a complementary topicalization effect (morality*topic interaction, $F(1, 278) = 3.89, p < .05$). This suggests that the source of the influence of morality on 'force' is likely instead in the interpretation of the ability component of its semantics.

4. Conclusion: Our approach builds on emerging research which understands modal flavors as arising from an interaction between the linguistic representation of modality and a psychological representation of possibility. We have described topicalization and order effects which are not well captured in a standard semantics for 'force' and 'able'. We have used those effects to come to a better understanding of the impact of the subjective dimension on judgments about 'force'. We believe our framework can be extended to help make sense of the impact of subjectivity on a variety of other judgments about apparently non-moral phenomena, helping us model the broad interaction of moral and modal judgments. Importantly, however, it will not extend to epistemic modals, if these have a standard semantics, since that semantics is not be sensitive to the posited causal background parameter. This rightly predicts an observed contrast between the strong effects of morality on judgments about agency, but not on corresponding judgments about epistemic modality. Our proposal thus moves us closer to a precise model of the influence of the subjective dimension on modal judgments.

References

Feltz, A. and Cokely, E. T. (2011). Individual differences in theory-of-mind judgments: Order effects and side effects. *Philosophical Psychology*, 24(3):343–355.

von Fintel, K. (1994). *Restrictions on Quantifier Domains*. PhD thesis, University of Massachusetts at Amherst.

von Fintel, K. (2001). Counterfactuals in a dynamic context. In *Ken Hale: A Life in Language*. MIT Press.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33(04):315–329.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1(337-355).

Kratzer, A. (1981). The notional category of modality. In Eikmeyer, H. and Rieser, H., editors, *Words, Worlds, and Contexts: New Approaches in Word Semantics*. de Gruyter.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1):339–359.

Mandelkern, M., Schultheis, G., and Boylan, D. (2015). I believe I can φ. In *20th Amsterdam Colloquium*, pages 256–265.

Mandelkern, M., Schultheis, G., and Boylan, D. (2016). Agentive modals. To appear, *The Philosophical Review*.

Phillips, J. and Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, 20(1):30–36.

Stalnaker, R. (1968). A theory of conditionals. In Rescher, N., editor, *Studies in Logical Theory*, pages 98–112. Oxford: Blackwell.

Young, L. and Phillips, J. (2011). The paradox of moral focus. *Cognition*, 119(2):166–178.