Proceedings of SALT 28: 000–000, 2018

Sticky situations: *Force* and quantifier domains*

Matthew Mandelkern
All Souls College, Oxford

Jonathan Phillips
Harvard University

Abstract When do we judge that someone was forced to do what they did? One relatively well-established finding is that subjects tend to judge that agents were not forced to do actions when those actions violate norms. A surprising discovery of Young & Phillips 2011 is that this effect seems to disappear when we frame the relevant 'force'-claim in the active rather than passive voice ($\ulcorner X$ forced Y to $\varphi\urcorner$ vs. $\ulcorner Y$ was forced to φ by $X\urcorner$). Young and Phillips found a similar contrast when the scenario itself shifts attention from Y (the forcee) to X (the forcer). We propose that these effects can be (at least partly) explained by way of the role of attention in the setting of quantifier domains which in turn play a role in the evaluation of 'force'-claims. We argue for this hypothesis by way of an experiment which shows that sequences of active vs. passive 'force'-claims display the characteristic "stickiness" of quantifier domain expansion, using a paradigm which we argue provides a useful general paradigm for testing quantifier domain hypotheses. Finally, we sketch a semantics for 'force' which we argue is suitable for capturing these effects.

Keywords: moral psychology, 'force', quantifier domains, experimental semantics and pragmatics, Knobe effect

1 Introduction

A wide variety of studies in recent years have shown that judgments about morality have surprising ramifications in apparently distinct conceptual domains, in particular having to do with causality and intentionality. Among other things, judgments about morality influence judgments about whether someone was forced to do what they did. In this paper we aim to improve our understanding of this phenomenon by making a case study of subtle variations in these judgments identified by Young & Phillips (2011). We argue that these judgments are best explained by way of a hypothesis which connects judgments about 'force' to quantification over a domain which we call a *causal background*: a set of propositions which represent the causal structure of a particular situation. The more causes we take into account, the more likely we are to judge that someone was forced to do what they did.

* Thanks to several anonymous referees and the audience at *SALT* 28; audiences at the University of Chicago and *SPP* 2017; and to Michael Glanzberg, Angelika Kratzer, and Ginger Schultheis, for very helpful comments and discussion.

©2018 Mandelkern and Phillips

2 The basic effect

Moral judgments influence intuitions about ostensibly non-moral issues in surprising, and now well-documented, ways (this is sometimes called the *Knobe effect* after Knobe 2003; see 2010 for a review). While the basic intuition behind these phenomena can be see at least as early as Aristotle (*NE* 1110^a8-9), an explosion of recent empirical work has shown that people's moral judgments influence their ordinary assessments of whether or not an agent acted intentionally (Knobe 2003; Leslie, Knobe & Cohen 2006; Cova 2016), whether an agent caused a given outcome (Alicke 1992; Knobe & Fraser 2008; Hitchcock & Knobe 2009), whether or not an agent acted freely (Phillips & Knobe 2009), and so on (Pettit & Knobe 2009; Phillips, Luguri & Knobe 2015).

Our focus here will be on the impact that moral judgments have on assessments of whether or not an agent was forced to do a given action. The first empirical work on this topic, Phillips & Knobe 2009, randomly assigned half of participants to read each of the two variants in the following vignette; the first variant involves a morally neutral action (throwing cargo overboard) while the second involves a morally bad action (throwing a person overboard).

(1) **[Cargo/Wife]** While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw his [wife's expensive cargo/wife] overboard.

Thinking quickly, the captain took [her cargo/his wife] and tossed [it/her] into the sea. While the [expensive cargo/captain's wife] sank to the bottom of the sea, the captain was able to survive the storm and returned home safely

After reading the vignette, participants in the cargo condition were asked whether they agreed or disagree with (2).

(2) *Cargo:* The captain was forced to throw his wife's cargo overboard.

Participants in the wife condition were likewise asked whether they agreed or disagree with (3).

(3) *Wife:* The captain was forced to throw his wife overboard.

Participants' judgments about whether the ship captain was forced diverged in the two cases. In the morally permissible (cargo) condition, participants judged much more that the captain was forced (to throw the cargo overboard) than in the morally impermissible (wife) condition, where participants were much less inclined to judge that the ship captain was forced (to throw the passengers overboard). This

phenomenon, in which people's assessments of whether someone was forced are impacted by their moral judgments about the action in question, has been replicated and extended in a number of different ways (see Young & Phillips 2011; Chakroff & Young 2015; Phillips et al. 2015).

A promising recent attempt to account for the influence of normative considerations in participants' assessments of 'force' (and more generally), which we will build on in what follows, is due to Knobe & Szabó 2013, who argue that we can capture the meaning of the sentences assessed by participants in these studies by way of a modal proxy. Thus (2) could be understood as contextually equivalent to (4), and (3) could be understood as contextually equivalent to (5).

(4) Given the circumstances, the captain had to throw the cargo overboard.

(5) Given the circumstances, the captain had to throw his wife overboard.

Knobe and Szabó provide evidence for this claim by replicating the earlier pattern of results using these proxy sentences. That is, they showed that participants agreed with (4) but not with (5), similar to the way in which they agreed with (2) but not (3). Knobe and Szabó then propose an account of the difference in participants' judgments by way of a semantics of modal auxiliaries, drawing largely on Kratzer 1977, 1981, but argue that the domain of possibilities quantified over by modal auxiliaries in these cases is of an "impure" flavor, meaning that deontic considerations partially determine which possibilities are included in the domain. In particular, they appeal to a principle they call *Hope*, which ensures that the relevant modal domain always contains a possibility in which no contextually salient norms are violated. As they point out, (5) will be false as long as there is some possibility in the domain in which the contextually salient norm against murder is not violated. By contrast, this principle will not affect the truth value of (4), since throwing cargo overboard does not violate a salient norm. Thus, assuming (3) is contextually equivalent to (5) and (2) is contextually equivalent to (4), Knobe and Szabó are able to derive the difference in force judgments.

3 Attention affects

We think that Knobe and Szabó's idea is promising, and we will build on the basic idea here, trying to refine it in what follows by giving a more explicit semantics for 'force' and providing independent evidence that judgments in these cases are indeed sensitive to quantifier domains. The starting point for our discussion is the striking discovery from Young & Phillips 2011 that the effect summarized in the last section is sensitive to whether subjects are presented with an active or a passive variant of the 'force' ascription. For example, Young and Phillips asked participants to read either the cargo or passenger variant of (6):

(6) While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw his [expensive cargo/passengers] overboard. Thinking quickly, the captain ordered one of his sailors to throw the [cargo/passengers] overboard. While the [cargo/passengers] sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

In this scenario, Young and Phillips replicated the basic effect of morality with questions like (7):

(7) Was the sailor forced to throw the [cargo/passengers] overboard?

That is, as expected, Young and Phillips found that subjects were more inclined to answer affirmatively to (7) in the cargo condition than in the passenger condition, just as one would expect given the findings summarized above. Surprisingly, however, Young and Phillips found no similar effect when they asked participants about (8):

(8) Did the captain force the sailor to throw the [cargo/passengers] overboard?

Rates of affirmative answers to (8) did not vary significantly across the two conditions (passengers vs. cargo). This is surprising, because (8) is just the active form of the passive question in (7), and, on the face of it, is intuitively equivalent to it.

It is not immediately clear how Knobe and Szabó's proposal can on its own account for this kind of difference. Their *Hope* principle is meant to be quite general, and thus one would think participants would be just as unwilling to agree that the captain forced the sailor to throw the passengers overboard as they are to agree that the sailor was forced to throw the passengers overboard by the captain; in both cases, one would expect *Hope* to require that the domain include possibilities in which the sailor does not violate the contextually salient norm against murder, and thus one would not expect participants to agree that the captain forced the sailor to throw the passengers overboard any more than they agree that the sailor was forced to throw the passengers overboard.

What then might explain this pattern? One possibility we find promising, following Young and Phillips' own interpretation of the data, is that the explanation does not have to do with the particular semantics of the passive, but rather with the fact that (8) draws attention to the captain more than (7) does. Support for this kind of explanation of this difference comes from another study reported by Young and Phillips (2011), in which they varied whether the vignette itself focused participants' attention on the role of the captain or on the role of the sailor, as in the two versions of (9).

(9) While he was sailing on the sea, a large storm came upon a [sailor/captain] on a ship. The waves began to grow larger, and the [sailor's/captain's] small vessel was too heavy. The [sailor's/captain's] ship would flood and the [sailor/captain] would drown if he didn't make it lighter. The only way that the [sailor/captain] could keep the ship from capsizing was to throw the passengers overboard. Thinking quickly, the captain of the ship ordered the sailor to throw the passengers overboard. The sailor threw the passengers overboard. While the passengers sank to the bottom of the sea, the [sailor/captain] was able to survive the storm and returned home safely.

Participants were observed to agree significantly more with (8) when their attention had specifically been drawn to the role of the captain rather than the sailor. These findings fit naturally with an approach to Young and Phillips' data on which the key factor affecting the contrast between the active and passive variations has to do with the change in attention these two constructions naturally bring out between the forcer and the forcee, respectively. In more detail, we will suggest that drawing attention to the captain's causal role in bringing it about that the sailor threw the [cargo/passengers] overboard makes it more likely that participants take the captain's causal influence into account in assessing whether the sailor was forced to act as he did.

4 Domain expansion

Why does calling attention to the *forcer* lead to increased agreement with 'force' sentences in the immoral (passenger) condition, while calling attention to the *forcee* leads to decreased agreement in that condition? This is the question that we will focus on in the rest of this paper. Our hypothesis is that attentional variation affects the setting of a quantifier domain which plays a part in the interpretation of 'force'.

This is a *prima facie* natural hypothesis given that, in general, drawing attention to things—objects, possibilities, actions—can change judgments about the truth-value of sentences that involve quantificational structures. This hypothesis is also in line with the general strategy pursued by Knobe and Szabó above, on which modal quantifier domains play a crucial role in explaining the observed judgments. Importantly, if this claim is right, then it will help resolve the paradox that Young and Phillips identify regarding the effect of attention on 'force' judgments, i.e., that participants agree with 'The captain forced the sailor to throw the passenger overboard' while disagreeing with 'The sailor was forced to throw the passenger overboard by the captain'. While this combination of attitudes appears inconsistent at first, if the meaning of these sentences varies with the domain made salient by context, then it may turn out that there is nothing truly inconsistent about this combination of judgments.

Let us begin by considering a few examples that illustrate in a general way the role of attention in quantifier domain expansion. One place that this kind of effect can be seen is in the domain of generalized quantifiers like 'all', 'some', 'no', and so on. Suppose Ted has just finished the last beer in his house; then (10) sounds reasonable.

(10) There's no more beer.

Here Ted is only talking about the beer in his house: the domain of his quantifier ranges only over the things in the house. But now suppose Mark responds with (11):

(11) Sure there is—at the liquor store!

With his response, Mark is calling attention to beer that Ted was (quite reasonably) neglecting. This response seems to *expand* the salient domain of quantification for 'no'. An important feature of cases like this, observed by Lewis (1979), is that once our attention has been called to the beer in the liquor store, it is no longer easy to ignore that beer. Ted can complain that Mark's response is pedantic, unhelpful, and obnoxious; but it seems difficult, after Mark's response, for Ted to stand his ground and insist that (10) is true.

Importantly for present purposes, we find similar effects in broadly modal domains. For instance, in the case of knowledge ascriptions—which, following Hintikka 1962, we can treat as universal quantifiers over a set of possible worlds—simply drawing attention to skeptical possibilities makes it harder to subsequently ignore them (and thus harder to subsequently ascribe knowledge; see DeRose 1991; Lewis 1996). Likewise, in the case of conditionals—which, following Stalnaker 1968; Lewis 1973; Kratzer 1981 are standardly treated as devices for talking about worlds in a given domain of accessible worlds—calling attention to certain outlying possibilities makes it harder to subsequently ignore them. A pair of sequences, from Lewis 1973 (attributed to J. Howard Sobel) and von Fintel 2001 (attributed to Irene Heim), illustrates this phenomenon nicely:

(12) a. If the USA threw its weapons into the sea tomorrow, there would be war.
b. But if the USA and the other nuclear powers all threw their weapons into the sea tomorrow, there would be peace.

(13) a. If the USA and the other nuclear powers all threw their weapons into the sea tomorrow, there would be peace.
b. ?? But if the USA threw its weapons into the sea tomorrow, there would be war.

In the first case, both conditionals sound true. But when the order is reversed in the second case, so that the outlying possibility of all powers eliminating nuclear

weapons is raised *first*, it is harder to then ignore that possibility and hear a true reading of the second conditional (13b).

A final example, which comes closest to our present topic, concerns judgments about ability and freedom. Hawthorne (2001) argues that judgments about ability and freedom are sensitive to domain shifts in similar ways to knowledge ascriptions. Suppose that John eats a bagel. He says:

(14) I was able to do otherwise than eat a bagel.

There are contexts in which we would naturally judge (14) to be true. But as more and more of the causal background of John's actions is made salient, our inclination to do so somehow diminishes. As Hawthorne puts it:

> Were I to have embarked on increasingly philosophical reflection, noting first the neurological springs of my action and then, indeed, that my act the laws of nature over which I had no control, coupled with the distribution of microparticles in the distant past over which I had no control,... I would then find myself in the position where I could no longer with good conscience ascribe to myself free will concerning whether or not I had a bagel. As I think harder about the range of determinants of my action, my inclination to ascribe freedom of choice to myself withers. (Hawthorne 2001: p. 66-67)

And, intuitively, this kind of domain expansion, once again, seems difficult to reverse: once we tend to the full range of an action's causal background, we are less inclined regard actions as having been done freely.

5 Operationalizing domain expansion

One feature of domain expansion which we find across all these different domains is its order asymmetry. Across these different domains, it appears to be easier to expand domains than to shrink them; once a thing or a possibility is made salient, it is difficult to subsequently ignore it. This characteristic signature of domain expansion makes it possible to test domain expansion hypotheses in a systematic way. At a general level, if A and A' mean the same thing except that A' expands a quantifier domain, then, in a sequence of assertions $\langle A, A' \rangle$, we predict that rates of agreement with A will differ from rates of agreement with A'. 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. This, in turn, means that we can test quantifier domain hypotheses by presenting subjects with sequences of this form, and checking whether we find the expected "sticky" signature of quantifier domain expansion. In

this section, we use this test to explore whether the variation in judgments of passive vs. active 'force' sentences can be explained by the expansion of a quantifier domain.

5.1 Methods

We tested our quantifier domain hypothesis about 'force' by exploring ordinary judgments in sequences of 'force'-claims like the following:

(15) $\langle passive, active \rangle$:

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

b. The captain forced the sailor to throw the [passengers/cargo] overboard.

(16) $\langle active, passive \rangle$:

a. The captain forced the sailor to throw the [passengers/cargo] overboard.

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

If the variation in judgments between active and passive forms is indeed due to quantifier domain expansion, then we should find evidence for the characteristic pattern of domain expansion just surveyed: flat judgments at a fixed level d in one order, and a shift in judgments in the other order, from a first judgment at a level different from d, to a subsequent judgment close to d.

The predictions, sample size, materials, and analysis plan were all pre-registered before data were collected (https://aspredicted.org/6qa2t.pdf). All of the data, materials, and analysis code are also available (https://osf.io/crxdq/). We aimed for a sample size of 800 and recruited 812 participants in total ($M_{age} = 36.65$, $SD_{age} = 11.89$; 483 females, 2 unreported) from Amazon Mechanical Turk (https://www.mturk.com) (Buhrmester, Kwang & Gosling 2011). Participant recruitment was automated through TurkPrime to prevent repeat participation and limit recruitment to participants with a previously established high approval rate. Participants were randomly assigned to read either the passengers or the cargo version of the following scenario:

(17) While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy and the ship would flood if he didn't make it lighter.

The only things on the captain's small boat were a single sailor, some small but expensive cargo that he was transporting, and a number of passengers. He knew he had to throw something overboard to keep the ship from capsizing. Thinking quickly, the captain ordered the sailor to throw

> the [passengers/cargo] overboard. While the [passengers/cargo] sank to the bottom of the sea, the captain and his ship survived the storm and returned home safely.

After reading, participants were sequentially asked to indicate their agreement with two 'force' statements, one passive and the other active—given either in ⟨*passive*, *active*⟩ order, as in (15), or in the ⟨*active*, *passive*⟩ order, as in (16). In all cases, participants responded on a scale from 0 ('Disagree') to 100 ('Agree'). They then completed a brief demographic form, and were thanked for their time.

5.2 Results and discussion

We did not exclude any participants from the analysis. Participants' agreement ratings were analyzed using the lme4 package in R (Bates, Maechler, Bolker, Walker et al. 2014) by comparing a series of linear mixed-effects models, with morality (passengers vs. cargo) and sentence sequence ((15) vs. (16)) as between-subjects fixed factors, and judgment-order (judged first vs. judged second) as a within-subjects fixed factor. The significance of the effects was calculated by comparing a model that included the term in question to a model that differed only in excluding that term (Barr, Levy, Scheepers & Tily 2013).

Using this approach, we observed the predicted three-way interaction between morality, sentence sequence, and judgment order $\chi^2(1) = 31.154$, $p < .001$ (Fig. (1)). We decomposed this by investigating the relationship between sentence sequence and judgment order when the passengers vs. cargo were thrown overboard.

When the passengers were thrown overboard, we observed the hypothesized interaction between sentence sequence and judgment order, $\chi^2(1) = 28.997$, $p < .001$. This interaction can be characterized as follows. In the sequence in (15), participants agreed less with the passive form (judged first) ($M = 43.92$, $SD = 36.2$) than with the active form (judged second) ($M = 59.5$, $SD = 34.0$), $t(395) = -4.427$, $p < .001$, $d = 0.444$. In the sequence in (16), by contrast, there was no significant difference in agreement with the active form (judged first) ($M = 61.1$, $SD = 35.5$) and the passive form (judged second) ($M = 60.9$, $SD = 35.2$), $t(399) = 0.061$, $p = .952$, $d = 0.006$.

When the cargo was thrown overboard, we observed a weaker interaction between sentence sequence and judgment order, $\chi^2(1) = 4.093$, $p = .043$. This interaction arose from a qualitatively different pattern. In the sequence in (15), participants agreed *more* with the passive form of the sentence (judged first) ($M = 76.4$, $SD = 26.9$) than with the active form of the sentence (judged second) ($M = 70.7$, $SD = 30.9$), $t(401) = 1.983$, $p = .048$, $d = 0.198$. In the sequence in (16), there was again no difference in participants agreement with the the active form (judged first)

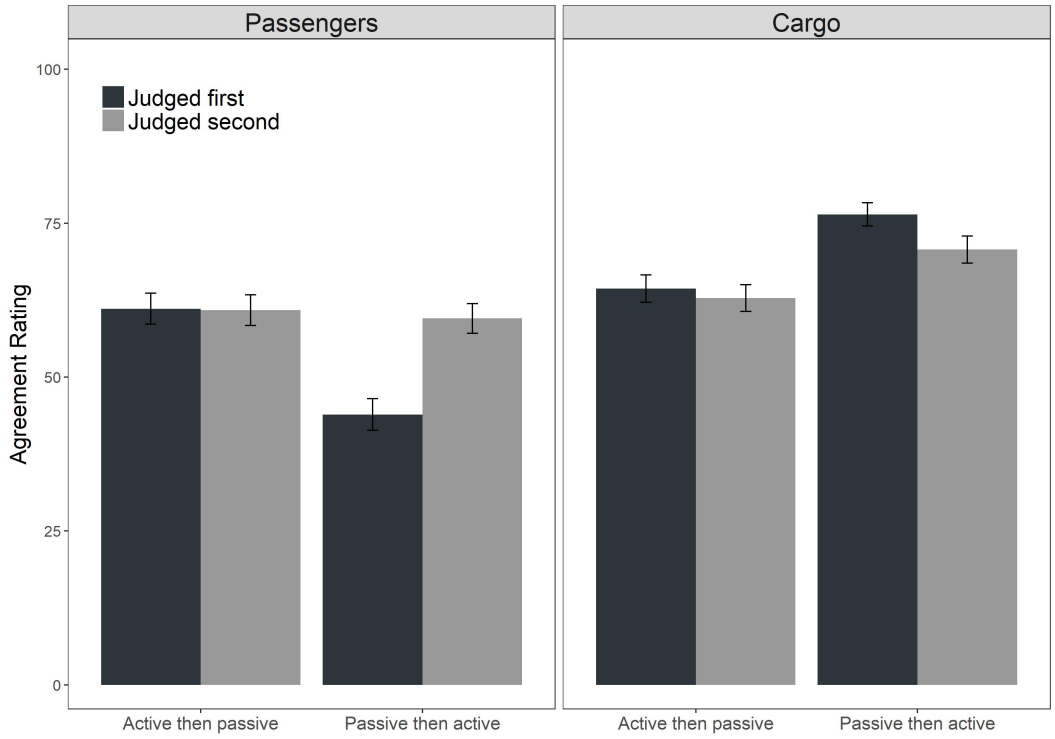

Figure 1 Mean agreement rating with force statements when the passengers (left panel) or cargo (right panel) were thrown overboard. Within each panel, agreement with the sentence sequence in (16) is on the left while agreement with the sequence in (15) is on the right. Color indicates whether the sentence was judged first (dark bars) or second (light bars). Error bars indicate $+/-1$ *SEM*.

($M = 64.4$, $SD = 31.8$) and the passive form (judged second) ($M = 62.8$, $SD = 31.0$), $t(406) = 0.493$, $p = .622$, $d = 0.049$.

Finally, to further characterize the observed effects, we investigated the direction of the shift in agreement with the passive sentences before vs. after reading the active sentences. In the passengers conditions, participants less agreed with the passive form of the sentence before reading the active sentence ($M = 43.92$, $SD = 36.2$) than after reading the active sentence ($M = 60.9$, $SD = 35.2$), $t(399) = -4.756$, $p < .001$, $d = 0.475$. In the cargo conditions, by contrast, participants agreed with the passive sentence *less* after reading the active sentence ($M = 62.8$, $SD = 31.0$) than before ($M = 76.4$, $SD = 26.9$), $t(397.15) = 4.720$, $p < .001$, $d = 0.468$.

What do these results show? Let us focus first on the passenger scenario. These results evidence exactly the kind of effect we would expect if a quantifier domain

expansion hypothesis were correct. Agreement with the 'force' sentences is lower when we start with the passive version, highlighting the sailor's contribution; it increases when the captain is subsequently attended to. By contrast, in the reverse, active-passive, order, agreement is relatively high when the captain is first attended to, and remains unchanged when we then attend to the sailor; in both cases it is at roughly the same point as in the active variant of the passive-active order. This evidences the "sticky" rates of agreement characteristic of quantifier domain expansion, with the higher agreement, active voice judgments (which focus attention on the captain) being the stickier ones—and thus the ones that go along with a larger domain of quantification.

Intriguingly, we don't find the same pattern in the cargo scenario. Here we have mid-point agreement with both sentences in the passive-active sequence. In the active-passive order, we have elevated agreement with the active sentence, and then slightly lower, but still elevated, agreement with the passive sentence. Given the relatively small difference between these judgments, it is not exactly clear how to interpret these data. If there were a clearer contrast across the passive-active sequence, so that the active variant was closer to the midpoint and thus to both variants in the active-passive sequence, we would have clearer evidence for a quantifier domain hypothesis here—but in the opposite direction as in the passenger case. As things are, however, the evidence in the cargo scenario is unclear.

6 The meaning of 'force'

In this section, we will take a step back from the experimental data presented so far and develop a semantics of 'force' which we will then argue argue can help us make sense of those data. In particular, we propose that the following provides a plausible semantics (we formulate this in the active voice, but we assume that, at a semantic level, the active and passive formulations are equivalent):

(18) $[\![\text{A forced B to } \varphi]\!]^{c,w} = 1$ 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 $[\![\varphi]\!]^c$'.

The first clause here ensures that 'force' is a success term: if Sue forced John to eat a cookie, then John must indeed have eaten a cookie. The second clause encodes the sense of *compulsion* characteristic of 'force' judgments: if Sue forced John to eat a cookie, then John must have been *compelled* to eat the cookie; there must be

a sense in which John was unable to do otherwise than eat the cookie. The third clause ensures that the forcer played the right kind of role in bringing about this compulsion: it is because of Sue that John was unable to do otherwise. The role of the final clause is to rule out Frankfurt-style cases involving 'force'. Suppose that Sue ensured that John could not do otherwise than eat a cookie, but John nonetheless ate a cookie completely of his own accord. In such a case, we do not want to say he was forced to eat a cookie. The fourth clause ensures that in situations like this, 'force' sentences will be false.

We think this provides a plausible first-pass account of the meaning of 'force', and provides a foothold in accounting for the effects of morality on judgments about force. We think that at least *one* source of the sensitivity of 'force' to morality and attention comes by way of the second component in the semantics here: 'force' entails a lack of ability, and judgments of morality and attention can affect ability ascriptions. Consider our core scenario. In the cargo condition, we are intuitively inclined to say that the sailor *could not have done otherwise than he did*; whereas in the passenger condition, we are much less inclined to say this. This suggests that morality has a similar effect on ability ascriptions as on force judgments: we are more inclined to judge agents to have been unable to do otherwise when what they do is morally neutral or good than when it is morally bad. This is exactly what we expect if 'force' has roughly the meaning we suggest above and if the influence of morality on force judgments comes (at least in part) by way of the ability component of the meaning of 'force'.

Moreover, the considerations Hawthorne raises provide independent intuitive evidence that attention can affect judgments about ability. Paying attention to more of the causal background of a given act can make us less likely to judge the agent as having been able to do otherwise. And in the passenger case, elevated 'force' judgments corresponded with highlighting the causal background of the sailor's act—by highlighting the captain's role in the sailor's act, through Young and Phillips' active voice formulation and attention manipulation—suggesting that this may be just the mechanism that is responsible for the attention effects observed with 'force'.

We will propose, then, that judgments about 'able' vary with a parameter which we will call a *causal background*. The causal background is a representation which keeps track of the causal structure of a given situation. There are different ways we could spell this out formally. We will use a simple model, in which causes are true propositions (intuitively, things like the denotation of 'The captain ordered the sailor to do such and such'), and the causal background is a set of propositions. In a given context, we may be happy to ignore certain features of the causal background and thereby represent agents as free to do other than they did. The more causes we take account of, however, the less we will be representing agents as free to do otherwise than they did—right up to Hawthorne's limiting case in which there is only one

possibility consistent with the causal background, namely the one in which the agent does exactly as they in fact did.

There are different ways we could integrate this kind of picture into a semantics for 'able'. The simplest approach would be to build on the standard Kratzerian picture (Kratzer 1977, 1981), and simply treat 'able' as a quantifier over the intersection of the causal background (or some subset thereof). But this approach has two drawbacks. One is a general concern that treating 'able' as an existential quantifier over worlds in this way gives us implausibly weak truth-conditions for ability ascriptions; see Mandelkern, Schultheis & Boylan 2017 for an extended recent discussion.

A more local concern has to do with the ways in which ability ascriptions are context sensitive. In our discussion above, we noted one way in which 'able' is sensitive to attention in context: drawing attention to an action's causal background can suppress agreement with ability judgments, in the way Hawthorne emphasizes. But there is another way in which ability ascriptions are context sensitive: as Mandelkern et al. (2017) demonstrates, drawing attention to certain *actions* can *increase* agreement with ability judgments. To see the point, suppose that Sue is planning to go to dinner with her friends. Ann says:

(19) Can you go to a movie tonight?

It's very natural for Sue to respond 'No, I can't; I have dinner plans'. But now suppose that Ann replies: 'So what? You *could* just cancel those plans and come to the movie. So you *can* go to the movie.' Once Ann has made this possibility salient, it becomes much harder for Sue to stand by her claim that she can't go. This case suggests that domain expansion in the case of ability ascriptions can go in two directions: while, as we saw above, calling attention to the *causal background* of some action can durably *depress* agreement with ability ascriptions, calling attention to certain *actions* available to an agent can durably *elevate* agreement with ability ascriptions. And moreover, intuitively, both kinds of change are sticky in the way characteristic of quantifier domain expansion. This suggests that there are two different roles for domains of quantification in ability judgments: one domain which represents the causal background and suppresses agreement with ability ascriptions as it grows; and one which represents contextually available actions and increases agreement with ability ascriptions as it grows.

We can capture this by building on the framework developed in Mandelkern et al. 2017. In that framework, ⌜S is able to φ⌝ is evaluated relative to a set of actions $\mathscr{A}$, and is true just in case there is some action A in $\mathscr{A}$, such that if S tried to do A, she would do φ. What set of actions $\mathscr{A}$ amounts to is determined by context (Mandelkern et al. call this the set of *practically available actions*); importantly, expanding that set of actions will *elevate* agreement with ability ascriptions, since it

will expand the possible paths for S to do φ that we are treating as salient in a given context. This account thus nicely captures the second kind of domain-sensitivity we explored just now. To capture the first kind, we can distinguish a separate parameter which represents the context's causal background, and which provides the building blocks for the practically available actions. More formally, let $\mathscr{C}$ denote the causal background—again, a set of propositions representing the salient causal structure of the situation in question. We can then think of $\bigcap\mathscr{C}$ as the building blocks for $\mathscr{A}$: $\mathscr{A}$ will be a set of propositions whose elements are all in $\bigcap\mathscr{C}$. Then, crucially, as $\mathscr{C}$ grows—as we take into account more of the causal background—$\mathscr{A}$ will correspondingly generally shrink, as we exclude from $\bigcap\mathscr{C}$ worlds that were in elements of $\mathscr{A}$. And so, while expanding the set of practically available actions will durably elevate agreement with ability ascriptions, expanding the causal background by attending to more of the causal structure of a given situation will durably depress agreement with ability ascriptions.

Putting these pieces together formally: where $\wp(\bigcap\mathscr{C}_{c,w})$ is the powerset of the intersection of the causal background $\mathscr{C}_{c,w}$ at a given context c and world w, and f_c is the contextual selection function from Stalnaker (1968)'s theory of conditionals, which takes a proposition and a world to the "closest" world where that proposition is true:

(20) $[\![\text{S is able to } \varphi]\!]^{c,w}$

a. is defined only if $\mathscr{A}_{c,w} \subseteq \wp(\bigcap\mathscr{C}_{c,w})$;

b. where defined, is true iff $\exists A \in \mathscr{A}_{c,w}$ such that $f_c(S \textit{ tries to do } A, w) \in [\![\text{S does } \varphi]\!]^{c,w}$

Informally, ⌜S is able to φ⌝ is true just in case there is some practically available action which entails everything in the causal background, and which is such that if S tries to do that action, she does φ.

Mandelkern et al. (2017) motivate the core truth conditions given in (20b), and we won't focus on them here. What is important for present purposes is that this approach captures the dual roles of domain expansion in ability ascriptions. On the one hand, attending to more causal background, as in Hawthorne's fatalistic reverie, will expand the causal background and thus shrink the set of practically available actions. This will reduce the number of possible ways we take into account when thinking about whether the agent could bring about φ—and thus durably diminishing agreement with ability ascriptions. In the limiting case—determinism—the causal background includes every causal factor that constrained the agent's action, resulting in it being the case that the only practically available action is what the agent actually did. On the other hand, directly attending to certain actions, as in Ann's reply 'So what? You *could* just cancel those plans and come to the movie', can expand the

range of actions we treat as practically available. This will increase the number of possible ways we take into account when thinking about whether the agent could bring about φ, and thereby durably increase agreement with ability ascriptions.

7 Explaining our data

With these semantics in hand for 'force' and 'able', let us revisit the experimental results discussed in the first part of the paper. First, to explain the basic effect of morality, we will assume that something along the lines of what Knobe and Szabó proposed is correct, appropriately generalized to the present framework. That is, we assume that subjects generally work with a causal background whose intersection includes possibilities in which people do the right thing—and thus with corresponding sets of actions which include actions which do not violate moral norms. This means that in cases like the passenger scenario above, subjects will be less likely to take into account the full causal background which led the sailor to throw the passengers overboard, and they will be more likely to include actions in the set of practically available actions in which the sailor does something other than what he actually did, than in the cargo scenario.

But this tendency to treat agents as free to do morally good acts can run against a countervailing tendency when we call attention to more of the causal background in a given situation. In particular, highlighting the captain's causal role in bringing about the sailor's action—for instance, by using the active voice, or highlighting the captain when describing the scenario—will tend to make us more likely to include in the causal background propositions that capture the captain's causal role: e.g., the propositions that the captain ordered the sailor to throw the passengers overboard, that the captain was a superior to the sailor, and so on. Given our semantics for 'able' above, this in turn will make us less likely to treat the sailor as able to have done otherwise than he in fact did. And, given our semantics for 'force', that, in turn, will make us *more* likely to say that the sailor was forced to do what he did: for on our semantics for 'force', 'The captain forced the sailor to throw the passengers overboard' can only be true if the sailor was *not* able to do otherwise than throw the passengers overboard. In short, decreasing agreement with 'The sailor was able to do otherwise than throw the passengers overboard' by expanding the causal background will increase agreement with its negation, and thus tend to increase agreement with the corresponding 'force'-claim.

This provides an account of the attention effects in the passenger case, and it also predicts the order effects we observed above. If quantifier domain expansion is sticky, then expanding the causal background will lead to *durably* lower agreement with ability ascriptions, and thus to durably higher agreement with force ascriptions. And this is just what we observe: if we first focus attention on the sailor and then

focus attention on the captain, we see an increase across the sequence in rates of agreement; whereas if we start by focusing attention on the captain and then subsequently focus on the sailor, we start from the beginning with relatively high rates of agreement in the 'force'-claim, and this stays stable across the sequence.

As we saw above, we do not find similar order effects in the cargo case. A natural explanation of this is that in the cargo case, we are less likely to be ignoring the captain's causal role in the first place, and so calling attention to the captain does not change judgments in the same way. As we noted above, there is variation in the judgments in the cargo case. They seem to be consistent with quantifier domain variation, but are nowhere near as clear as the pattern in the passenger case, and go in the opposite direction. We do not see a natural way to tell a quantifier domain story about these judgments. Another possibility is that the differences in judgments observed here has to do with independent factors having to do with the questions made salient by active vs. passive formulations and the meaning of 'force', but not to do with quantifier domains. More work is needed on this topic.

8 Conclusion

It is now well-known that morality exerts a surprising influence on judgments about ostensibly non-moral domains such as intentionality, causality, and compulsion. In this paper we sought to make progress on understanding morality's influence by exploring subtle judgments about whether an agent was forced to do what they did. In particular, we have focused on Young & Phillips (2011)'s surprising finding that variation in whether the 'force'-claims in question are formulated in the active or passive voice modulates morality's impact. We have argued that this modulation has to do with the influence of attention—which can be affected by active/passive variation, as well as by manipulating who the scenario focuses on—on a quantifier domain which affects the interpretation of 'force'. We presented an experiment which showed that sequences of active versus passive 'force' judgments show exactly the classic sticky pattern we expect of quantifier domain expansion. We argued that the interpretation of 'force' depends in part on a *causal background* which keeps track of the causal structure of a given scenario; the more causes we take into account, the more we will judge someone to have been forced to do what they did. This provides us with the resources to make sense of the effect of attention on 'force' judgments in immoral cases.

In concluding, we want to step back and emphasize some of the more general upshots of our exploration. One concerns moral psychology. How should we think about the participants who show are less inclined to accept 'The sailor was forced to throw the passengers overboard by the captain' than 'The captain forced the sailor to throw the passengers overboard'? One perspective, which Young and Phillips

embrace, is that this inclination amounts to an 'internal error', landing us amidst 'logical inconsistencies'. We have suggested a different way of thinking about this inclination. On our proposal, the results in question do not point to logical inconsistencies in participants' judgments. Rather, participants in the two conditions—active vs. passive voice—are evaluating different contents, made prominent by changes in the information structure of the sentences which affect context-sensitive terms like 'force', and there is nothing inconsistent about having different attitudes towards those different contents. On this alternative perspective, it would be wrong to regard the observed sensitivity to morality and attention as fundamentally erroneous.

Of course, we might still conclude that the observed sensitivity to morality and attention is rationally suboptimal. We could try to imagine an agent whose concepts of intsentional action, causation, or compulsion were not sensitive to morality or attention. Perhaps such an agent would have one up on us rationality-wise. But perhaps not: the forms of moral context-sensitivity we have observed in natural language may turn out to be perfectly rational once one takes into account humans' fundamentally social nature and their limited capacity to represent and reason over non-actual possibilities (Phillips & Knobe 2018; Phillips & Cushman 2017; Shtulman & Phillips 2018). This is a central question which moral psychology must grapple with.

A second upshot is for experimental semantics and pragmatics. It is difficult to find grounds to argue that some set of judgments is a feature of context sensitivity rather than simple inconsistency or irrationality. In this paper, we have operationalized a well-known intuitive observation about quantifier domain expansion—its stickiness—showing that this feature can provide the foundations for experimental investigation of whether a given phenomenon is indeed due to changes in some domain of quantification. Specifically, the stickiness of quantifier domain expansion provides a characteristic signature in rates of agreement across sequences of the form $\langle A, A' \rangle$ and $\langle A', A \rangle$, respectively. Investigating sequences of this form should provide a solid foundation for claims about quantifier domain expansion well beyond the present inquiry. This, in turn, raises an important theoretical question which, to our knowledge, remains open, namely *why* quantifier domains are sticky in this way. This question is important for understanding why the present paradigm works, and, we suspect, for better understanding the nature and sources of context-sensitivity and the cognitive effects of attention in general; it is just one of many important questions we leave open here.

References

Alicke, Mark D. 1992. Culpable causation. *Journal of personality and social psychology* 63(3). 368.

Barr, Dale J, Roger Levy, Christoph Scheepers & Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68(3). 255–278.

Bates, Douglas, Martin Maechler, Ben Bolker, Steven Walker et al. 2014. lme4: Linear mixed-effects models using eigen and s4. *R package version* 1(7). 1–23.

Buhrmester, Michael, Tracy Kwang & Samuel D Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6(1). 3–5.

Chakroff, Alek & Liane Young. 2015. Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition* 136. 30–37.

Cova, Florian. 2016. *The folk concept of intentional action* 117–141. John Wiley & Sons, Ltd. doi:10.1002/9781118661666.ch8.

DeRose, Keith. 1991. Epistemic possibilities. *Philosophical Review* 100(4). 581–605.

von Fintel, Kai. 2001. Counterfactuals in a dynamic context. In *Ken Hale: A life in language*, MIT Press.

Hawthorne, John. 2001. Freedom in context. *Philosophical Studies* 104(1). 63–79.

Hintikka, Jaakko. 1962. *Knowledge and belief: An introduction to the logic of two notions*. Ithaca, NY: Cornell University Press.

Hitchcock, Christopher & Joshua Knobe. 2009. Cause and norm. *The Journal of Philosophy* 106(11). 587–612.

Knobe, Joshua. 2003. Intentional action and side effects in ordinary language. *Analysis* 63(279). 190–194.

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

Knobe, Joshua & Ben Fraser. 2008. Causal judgment and moral judgment: Two experiments. *Moral psychology* 2. 441–8.

Knobe, Joshua & Zoltán Gendler Szabó. 2013. Modals with a taste of the deontic. *Semantics and Pragmatics* 6(1). 1–42.

Kratzer, Angelika. 1977. What 'must' and 'can' must and can mean. *Linguistics and Philosophy* 1(3). 337–355.

Kratzer, Angelika. 1981. The notional category of modality. In H. Eikmeyer & H. Rieser (eds.), *Words, worlds, and contexts: New approaches in word semantics*, de Gruyter.

Leslie, Alan M, Joshua Knobe & Adam Cohen. 2006. Acting intentionally and the side-effect effect theory of mind and moral judgment. *Psychological Science* 17(5). 421–427.

Lewis, David. 1973. *Counterfactuals*. Oxford: Blackwell.

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

Lewis, David. 1996. Elusive knowledge. *Australasian Journal of Philosophy* 74(4). 549–567.

Mandelkern, Matthew, Ginger Schultheis & David Boylan. 2017. Agentive modals. *The Philosophical Review* 126(3). 301–343.

Pettit, Dean & Joshua Knobe. 2009. The pervasive impact of moral judgment. *Mind & Language* 24(5). 586–604.

Phillips, Jonathan & Fiery Cushman. 2017. Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences* 114(18). 4649–4654. doi:10.1073/pnas.1619717114. http://www.pnas.org/content/114/18/4649.abstract.

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

Phillips, Jonathan & Joshua Knobe. 2018. The psychological representation of modality. *Mind & Language* 33(1). 65–94. doi:10.1111/mila.12165. http://dx.doi.org/10.1111/mila.12165.

Phillips, Jonathan, Jamie B. Luguri & Joshua Knobe. 2015. Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition* 145(Supplement C). 30 – 42. doi:https://doi.org/10.1016/j.cognition.2015.08.001.

Shtulman, Andrew & Jonathan Phillips. 2018. Differentiating 'could' from 'should': Developmental changes in modal cognition. *Journal of Experimental Child Psychology* 165. 161–182. doi:https://doi.org/10.1016/j.jecp.2017.05.012. http://www.sciencedirect.com/science/article/pii/S0022096517303429.

Stalnaker, Robert. 1968. A theory of conditionals. In Nicholas Rescher (ed.), *Studies in logical theory*, 98–112. Oxford: Blackwell.

Young, Liane & Jonathan Phillips. 2011. The paradox of moral focus. *Cognition* 119(2). 166–178.

Matthew Mandelkern
All Souls College
Oxford, OX1 4AL
United Kingdom
matthew.mandelkern@all-souls.ox.ac.uk

Jonathan Phillips
Harvard University
1482 William James Hall
33 Kirkland Street
Cambridge, MA 02138
USA
phillips01@g.harvard.edu