# IMPURE MODALS[1]

Joshua Knobe and Zoltán Gendler Szabó
Yale University

Abstract: The aim of this paper is to present an explanation for the impact of normative considerations on people's assessment of certain seemingly purely descriptive matters. The explanation is based on two main claims. First, a large category of expressions are tacitly modal: they are contextually equivalent to modal proxies. Second, natural language modals can have impure flavors, being evaluated against conversational backgrounds shaped by heterogeneous considerations, including normative ones. This impurity in the conversational backgrounds can be modeled using a pair of principles which jointly make certain possibilities relevant at the expense of others.

When we assess an action as right or wrong we often want to know whether it was forced upon the agent, whether it had certain causal effects, and whether it was performed intentionally. It seems natural to think that these questions are prior to and independent of normative considerations. Recent experimental results have shown that in fact this is not so: when people consider such questions their answers often depend on whether the action has violated a norm.

Our aim here is to present a new account of these puzzling phenomena. It is tempting to see them as manifestations of diverse errors but we will argue that the data permit a unified and charitable explanation. The key is to abandon the assumption that all modals can be categorized into pure flavors: circumstantial, deontic, teleological, etc. Hence our title – we claim that some modality is impure.

## 1. The experiments

Before starting in with the explanation, it will be necessary to get a sense for the data. We begin with a brief summary of the three effects we want to focus on.

[1] We are grateful for comments from Michael Bratman, Tad Brennan, Mark Crimmins, Josh Dever, Kai von Fintel, Tamar Gendler, Shelly Kagan, Angelika Kratzer, Jonathan Phillips, Rob Rupert, Jonathan Schaffer, Stewart Shapiro, Ted Sider, Jason Stanley, Brian Weatherson and Seth Yalcin, as well as to audiences at the Arché Research Centre at the University of St. Andrews, the University of Cologne, the University of Colorado, Cornell University, Stanford University and Yale University.

### 1.1. Freedom

Consider whether normative considerations play a role in how people distinguish between actions an agent chooses freely and actions that are forced upon her. The standard way of testing such matters is to have people read brief vignettes and then answer questions about the actions of the characters in these vignettes. We can then systematically vary specific factors within the vignettes and thereby determine the role they play in shaping judgments. Thus, if we want to know whether normative considerations are having any impact, we can construct a pair of vignettes that are identical except for the normative status of the agent's behavior and then check to see whether we find a corresponding difference in the responses to the scenario described.

In one recent study (Phillips and Knobe 2009), participants were randomly assigned to one of two conditions. Participants in the first condition were given the following vignette:

> 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 overboard.
>
> Thinking quickly, the captain took her cargo and tossed it into the sea. While the expensive cargo sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

These participants were asked whether they agreed or disagreed with (1a).

(1a) The captain was forced to throw his wife's cargo overboard.

Participants in the other group received a vignette that was identical, except for changes designed to alter the normative status of the agent's behavior (changes shown in italics):

> 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* overboard.
>
> Thinking quickly, the captain took his *wife* and tossed her into the sea. While the captain's *wife* sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

These participants were asked whether they agreed or disagreed with (1b).

(1b) The captain was forced to throw his wife overboard.

The results showed a significant difference between conditions. Participants tended to disagree with (1b) but to agree with (1a). This difference provides at least some initial indication that normative considerations are playing a role in determining people's judgments about whether an agent has been forced to perform an action.

But, of course, a single experiment like this one can never provide decisive evidence. It will always turn out that the two vignettes differ in numerous respects – some of which have nothing to do with normative considerations *per se* – and one might always worry that one of these other differences is actually at the root of the observed effect. As long as one is relying just on a single experiment, it will be difficult to properly address this worry. The best approach, then, is to construct a number of different pairs of vignettes that share the same basic structure but that differ radically in their details (Phillips and Knobe 2009; Phillips and Young 2010). So, for example, in a separate study, participants received a vignette about a doctor who is ordered by the chief of surgery to prescribe a medicine that will either help a patient (in one condition) or harm a patient (in the other). The doctor is described as reluctantly agreeing to prescribe the medicine in both cases, but participants tend to say that he was forced in the help condition but not in the harm condition (Phillips and Knobe 2009). As we accumulate more and more pairs of vignettes that show this same pattern of responses, it begins to seem increasingly implausible to search for a separate explanation for the effect on each pair. The more parsimonious explanation is that people's normative judgments actually are having an impact on their judgments as to whether or not the agent was forced to act.

### 1.2. Causation

If the effect described in the previous section arose only for this one expression, it would be natural to suppose that it was due to some idiosyncratic feature of the verb 'force'. However, experimental results indicate that a similar effect arises for numerous other expressions. For example, one can find the same basic asymmetry in people's judgments about causation.

In one recent study of this effect (Knobe and Fraser 2008), all participants received the following vignette:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so

> do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

The actions of the two agents are similar: both of them take a pen, and if either of them had not done so, the problem would not have arisen. But the two actions differ normatively: only one of the agents is violating a rule. The key question now is whether this normative difference has an impact on people's causal judgments.

To get at this question, participants were asked whether they agreed or disagreed with (2a) and (2b).

(2) a. Professor Smith caused the problem.
b. The administrative assistant caused the problem.

The results showed significant difference. Participants tended to agree with (2a) while disagreeing with (2b). In other words, the normative difference between the two actions apparently had an impact on people's judgments about whether each of them counted as a cause.

This effect, too, has been replicated and extended in a number of subsequent studies. Such studies have shown that the effect continues to arise when one controls more carefully for purely statistical differences between events (Roxborough and Cumby 2009; Sytsma, Livengood and Rose 2010), when the normative considerations come from moral judgments rather than from a conventional rule (Cushman, Knobe and Sinnott-Armstrong 2009), when the word 'caused' is replaced with a lexical causative (Cushman et al. 2009), and even when the outcome itself is something good (Hitchcock and Knobe 2009). While considerable controversy remains about how exactly to explain these results, the impact of normative considerations on judgments about causation appears well-established.

### 1.3. Intentional Action

Finally, consider intentional action. One might initially suppose that the question as to whether or not an agent performed a behavior intentionally is an entirely descriptive question, simply a

matter of what the agent's intentions are and how the agent acts. Yet, once again, a series of experimental studies indicate that normative considerations can play a role.

To see the basic effect at work in such cases, consider the following vignette:

> Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bulls-eye. He raises the rifle, gets the bull's eye in the sights, and presses the trigger.
>
> But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…
>
> Nonetheless, the bullet lands directly on the bull's-eye. Jake wins the contest.

Now, keeping this vignette in mind, ask yourself whether you agree with the sentence:

(3a) Jake hit the bull's eye intentionally.

The experimental results indicate that most participants do not agree with this sentence (Knobe 2003). The issue here, presumably, is that the agent is succeeding entirely by luck; he doesn't really have any control over the outcome of his behavior.

But now suppose that we switch around the normative significance of the story. We can leave intact all of the facts about the process, while simply altering the target.

> Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in the sights, and presses the trigger.
>
> But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…
>
> Nonetheless, the bullet hits her directly in the heart. She dies instantly.

The corresponding sentence is then:

(3b) Jake hit his aunt intentionally.

Yet people tend to say that this latter sentence is true (Knobe 2003). Of course, one might initially suppose that the difference here arises simply because it is easier to hit an aunt than a bull's eye (Gugliemo & Malle 2010), but more tightly controlled studies show that the effect continues to arise even when the difficulty of hitting the two targets is kept the same (Sousa &

Holbrook 2010). The effect has also been shown to arise even when the agent is doing something that requires no skill of any kind, such as rolling some dice (Nadelhoffer 2004) or guessing a random number (Nadelhoffer 2005). In all of these cases, the experimental results indicate that people are more willing to consider the behavior intentional when it violates a norm than when it does not.

### 1.4. Two desiderata

Thus far, we have been reviewing three different effects from the experimental literature. Stating the results as neutrally as possible what we found is this. In each case, we consider a sentence containing a word that appears to be entirely descriptive, but we find that normative considerations influence how the word is used. The question now is how these various effects are to be explained.

One might seek separate explanations for separate effects but this sort of approach risks missing an insight into the pattern as a whole. It is reasonable to expect that the root of the unexpected impact of the normative is the same across the board. So, we place our bets on the opposite strategy: we believe the ideal explanation of these asymmetries should be a *unified* one. This is our first desideratum.

One way of providing a unified explanation would be to posit a performance error of some sort that impacts people's judgments in a number of different domains. That is, one could say that normative considerations actually are not relevant to any of these questions but that people are messing up in some way and allowing their judgments to be improperly influenced. Again, we think this approach should only be followed if all else fails: the ideal explanation for the encroachment should remain *charitable.* This is our second desideratum.

We expect that our second desideratum will raise more eyebrows than the first, so let us say a few words in its support. The first thing to note here is that it is not *obvious* that there has to be a performance error at work in these effects. Participants show many of the asymmetries even in 'within-subject' designs where they receive the two cases back-to-back and can clearly see that the only difference between them is a normative one (Knobe and Fraser 2008; Young et al. 2006). So one cannot simply dismiss these effects as some minor slip-up that people would quickly correct on further reflection.

But, of course, that is not the true test of the performance error approach. What we really want to know is whether it is possible to develop a specific performance error hypothesis that can accurately predict and explain the data. A number of researchers have provided possible hypotheses (Adams and Steadman 2004; Alicke 2008; Nadelhoffer 2006; Nanay 2010). Each of these hypotheses describes a specific error and explains how such an error could lead to the asymmetries observed. These hypotheses also make new predictions, which have been put to the test in further experiments.

There have been studies using reaction time measures (Gugliemlo and Malle 2011), patients with lesions to the specific brain regions (Young et al. 2006), participants with Asperger's syndrome (Zalla forthcoming), and numerous studies that simply use additional vignettes (Nichols and Ulatowski 2007; Sripada and Konrath forthcoming). In each case, the data have failed to vindicate the performance error hypotheses. Instead, studies have again and again shown that the predictions made by these hypotheses are not borne out. (For a review, see Knobe 2010.)

A possible response would be to say that although the existing performance error hypotheses happen not to be quite right, we simply have not yet picked out the right sort of error. This might ultimately prove correct, but our hunch is that it would be best at this point to begin looking elsewhere for a solution. Perhaps the reason we can't find the error in people's judgments in these cases is that there is no error to be found.

Charity towards the majority should not be combined with dismissing the minority. We will defend the view that normative considerations can and normally do impact our judgments about freedom, causation, and intentionality. But it is not part of our view that such an impact is inevitable. We can and sometimes do find ourselves in special contexts where conversational participants seek to diminish the effect of this impact. When we make a conscious effort to assess a situation with full objectivity and impartiality, perhaps we can diminish or entirely neutralize the effect shown in the examples above. Perhaps this is what happens in scientific or philosophical inquiry. Moreover, some people might be especially prone to think in this sort of way, and some of them might be inclined to stick with it even in everyday settings. A good explanation should leave room for this, and we will certainly try.

## 2. Modality as the common core

Our hypothesis is that the judgments under consideration are tacitly modal. Although there are important differences among the target sentences, we propose that in understanding each within the contexts of its respective vignette people exercise a certain modal assessment. It is this modal assessment, we suggest, that explains the impact of normative considerations.

At this point, the straightforward way to proceed would be to present and justify a modal semantics for 'force,' 'cause,' and 'intentionally'. With such a semantics at hand, we could say that the relevant judgments are tacitly modal simply because they are judgments of the truth or falsity of modal sentences. We will, however, opt for a less direct connection between our target sentences and modality. The most obvious reason for this is that we simply do not know what an adequate semantics of these words would look like. We do have some ideas, enough for an outline of the semantic contribution these words make to the simplest sentences in which they appear. But these ideas are controversial, and if we made an attempt to extend the proposal to even slightly more complex cases, we would need to defend a crushing number of semantic assumptions. Not only would this be distracting, the end result itself might be less than fully illuminating. For what we seek is the commonality in the contextual behavior of these expressions and this commonality would likely be obscured by all the differences a semantic analysis is bound to bring to the fore.

Our strategy in this section will be to associate with our target sentences overtly modal ones. We will claim that by and large people consider the modal proxies as equivalent with the original sentence (or its negation) within the contexts created by the vignettes. The aim is to reduce the three explanatory tasks to a single one: why normative considerations have an impact on seemingly non-normative modal claims.

### 2.1. From the Target Sentences to Modal Proxies

The significance of modality is clearest in the case of 'force'. Speakers who judge that the captain was forced to throw the cargo overboard are likely to justify their claim by asserting (1a), while those who deny that the captain was forced to throw his wife overboard will be likely to deny (1b′).

(1′) a. Given the circumstances, the captain had to throw the cargo overboard.

b. Given the circumstances, the captain had to throw his wife overboard.

Indeed, even those who disagree with the dominant judgments about these cases would probably not object to casting the disagreement in these modal terms. It appears that the judgments made in this case have modal correlates and that – at least in the context established by the vignettes – people feel comfortable moving back and forth between the original and the proxy.

To put this claim to the test, we conducted a new study. Participants received the very same stories about the captain and the storm that had been used in earlier research, but instead of being given the original sentences about whether the agent was forced, they were given the modal proxies (1a′) and (1b′).[2] The results showed that the asymmetry observed for the target sentences also arose for the modal proxies. Participants tended to agree with the claim that the captain had to throw the cargo overboard (1a′), but they tended to disagree with the claim that the captain had to throw his wife overboard (1b′).[3]

Turning to judgments about 'cause,' matters become a little bit more complex. A broad array of different researchers in both philosophy and psychology have suggested that causal judgments might in some way be connected to modal reasoning, but different researchers have offered interestingly different views about precisely how this connection works (Halpern and Pearl 2005, Hitchcock 2007, Lewis 1973, 2000, Woodward 2003). Our approach here will rely on the traditional claim that causes necessitate their effects. Thus, we suggest that when people agree with the claim that the professor caused the problem, they will also agree with the explicitly modal proxy (2a′), and when they disagree with the claim that the administrative assistant caused the problem, they will also disagree with the explicitly modal proxy (2b′).

[2] Participants were 42 people recruited through Amazon's Mechanical Turk. Each participant rated the sentence on a scale from 1 ('disagree') to 7 ('agree'). In addition, each participant was asked an explicitly moral question about which option would have been better on the whole for the captain to choose: throwing the cargo/wife overboard or not throwing the cargo/wife overboard.

[3] Ratings for the cargo condition ($M$ = 6.8, $SD$ = .4) were significantly higher than those for the wife condition ($M$ = 2.6, $SD$ = 2.3), $t$ (40) = 7.9, $p < .01$. We used mediational analysis to examine the relationship between condition, moral judgment and modal judgment. Condition had a significant impact on moral judgment ($\beta = .75$, $p < .01$). When moral judgment was entered as a regressor, the impact of condition on modal judgment decreased from $\beta = .78$, $p < .01$ to $\beta = .35$, $p < .01$. A Sobel test showed that this reduction was significant, $Z = 21.4$, $p < .01$. In other words, the difference between conditions appears to be impacting people's modal judgments in part by impacting their moral judgments.

(2′) a. Given the actions of the professor, the problem had to occur.

b. Given the actions of the administrative assistant, the problem had to occur.

To confirm this suggestion, we ran an additional experiment. Participants were given the story of the missing pens and then randomly assigned to evaluate a sentence about that vignette. Some participants were given one of the original causal sentences: 'The professor/administrative assistant caused the problem.' Others were given one of the proposed modal proxies (2a′) and (2b′).[4] Unsurprisingly, we replicated the original effect whereby people agree more with the causal sentence about the professor than they do with the causal sentence about the administrative assistant.[5] But we also found the same effect for the modal proxies: people showed a moderate level of agreement with the explicitly modal sentence about the professor but disagreed with the explicitly modal sentence about the administrative assistant.[6]

Turn now to the case of 'intentionally.' We will not be concerned here with the larger questions about what it means for an act to be intentional. Instead, we will be focusing only on the pattern of people's judgments in the particular cases described above. In these cases, one might justify the claim that the shooting of the target was not intentional by saying: 'That only happened through sheer luck.' Our aim now is to spell out this intuitive notion in more precise terms.

Let us begin by using the word *fluke* to pick out the lucky events that happened to occur in this case (e.g., the gun slipping in exactly the right way). The claim now is that the outcome was due to this fluke and not to the agent's mental states (e.g., an intention to hit the target). One might spell out this claim more precisely as (3′).

(3′) a. Given the fluke, Jake had to hit the target.

b. Given the fluke, Jake had to hit his aunt.

[4] Participants were 80 people recruited through Amazon's Mechanical Turk. Each participant rated the sentence on a scale from 1 ('disagree') to 7 ('agree'). Data were analyzed using a 2 (sentence type: causal vs. modal) x 2 (agent: professor vs. assistant) ANOVA. There was a significant main effect of agent, $F$ (1, 76) = 43.6, $p < .001$. There was no significant main effect of sentence type and no significant interaction.

[5] Mean rating for the causal statement about the professor: 5.4, mean rating for the causal statement about the assistant: 2.3, $t$ (33) = 6.6, $p < .001$.

[6] Mean rating for the modal statement about the professor: 4.5, mean rating for the modal statement about the assistant: 2.7, $t$ (43) = 3.5, $p = .001$.

Note that these final modal proxies differ from the others in an important respect. The claim here is that to the extent that the modal proxy is regarded as true, people will take the target sentence to be false. Thus, we will say that while (1a′), (1b′), (2a′) and (2b′) are *positive* modal proxies, (3a′) and (3b′) are *negative* ones. In other words, it is really the negation of the modal proxy that corresponds to the target sentence. (These modal proxies also differ from the others in that they make use of a stipulation that is hard to make precise. One could say that the fluke is the particular event that occurred when Jake's hand slipped on the barrel but the gun somehow stayed on target, but thinking about just what *that* event might be will quickly lead to puzzles of event individuation. The puzzles are made worse by the fact that the event was *random.* Such puzzles may easily make the reflective weary about passing judgments about (3a′) and (3b′) and this fact renders them less amenable to direct experimental test.)

Here is the summary of the target sentences and their proposed modal proxies:[7]

(1) The captain was forced to throw the (a) cargo/(b) his wife overboard.
(1') Given the circumstances, the captain had to throw (a) the cargo/(b) his wife overboard.

(2) (a) The professor/(b) The assistant caused the problem.
(2') Given the action of (a) the professor/(b) the assistant, the problem had to occur.

(3) Jake hit (a) the target/(b) his aunt intentionally.
(3') Given the fluke, Jake had to hit (a) the target/(b) his aunt.

Our plan is to explain the impact of normative considerations on people's judgments in the experiments by reducing it to an impact on their modal proxies, and then arguing that the latter is to be expected, given an independently plausible account of the interpretation of modals. This brings out what we take to be the common core of the phenomena under consideration and it also reduces the dialectical burden on the semantic component of our explanation. Contextualism about freedom, causation, or intentionality are controversial doctrines. But when it comes to modality, contextualism is the established view.[8]

[7] (1), (2), and (3) differ slightly from the sentences used in the original experimental studies. The changes are insubstantial; we made some simplifications to reduce verbiage.
[8] Setting aside epistemic modality for the moment, where relativist accounts are a strong competitor (Egan, Hawthorne, and Weatherson 2005, Stephenson 2007, but see Yalcin forthcoming). Portner 2009 is a standard survey of current views on the semantics of modals in natural languages; it mentions a wide variety of approaches, not one

For definiteness, we have offered a specific modal proxy for each of the original sentences whose use is to be explained. However, the core idea of the proposal does not depend on these precise modal correlates being the right ones. If you think that another, slightly different modal proxy might be more accurate, please hold on until Section 5. Then you can check to see whether the explanation we propose there works for your favored modal proxy as well.

### 2.2. The Nature of Modal Proxies

To begin with, we must clarify what exactly we take the relationship between (1) – (3) and their modal proxies (1′) – (3′) to be. This has to be stated with some care because the proxies certainly express different propositions in different contexts, and quite possibly the same holds of the target sentences as well. The tightest relationship would be mutual entailment in all contexts, but this is definitely too strong. To take one especially obvious example, suppose we introduce a modified version of the story of the captain and the storm. The boat is being menaced by a storm, but before the captain even notices the storm, he decides to throw his wife's expensive cargo overboard simply out of spite. Then it is intuitively still true that given the circumstances he had to throw the cargo overboard but it is certainly false that he was forced to do so. Finding counterexamples to the equivalence between the original sentence and its modal proxy is equally straightforward in the other cases.

Fortunately, what we need to get our explanation going is not mutual entailment in all contexts – it is merely mutual entailment in the particular context established by the vignettes. At the time when we are evaluating (1a) we take a lot of information about the case for granted. Some of this information has been conveyed *explicitly*: it is entailed by the text of the vignette. This is how we know that there was a storm threatening a ship, that the captain of the ship threw his wife's cargo overboard, and that the cargo sank to the bottom of the sea. Some other information has been conveyed *implicitly*. The most striking example is the proposition that the ship did not sink. We are told that the captain did what he could to save the boat, that he was able to survive the storm,

---

of which is invariantist. Most semantic approaches interpret modals via quantification over possibilities, and at the very least the domain of this quantification is supposed to be contextually determined. Many approaches to modal semantics employ an accessibility relation instead of explicit domain restriction, but these too accept that the relevant accessibility relation must be contextually provided. Semantic minimalists reject this sort of context-sensitivity; but they also part with mainstream semantics in rejecting that 'ready', 'tall', or 'usually' are context-sensitive. We reject semantic minimalism; for arguments see Szabó (2006).

and that he returned home safely, but all this is fully compatible with the possibility that the ship went down despite the captain's action, that he survived the storm on open sea, and that he was eventually rescued by another ship. Still, normal interpreters of the vignette will take it for granted that the ship was saved. Now consider what information is semantically encoded in (1a) and (1a′). Arguably, the former entails that the captain's action was influenced by something beyond his control of which he was aware, while the latter does not. We suggest that the vignette implicitly conveys the proposition that the captain's action was influenced by something of which he was aware but which was entirely beyond his control.

Our hypothesis is that whatever the truth-conditional differences between our target sentences and their modal proxies (or, in the case of negative correlation, between the target sentences and the negations of their modal proxies) is *canceled out* by the information conveyed by the vignette in which they are evaluated. Let's say that two sentences are *contextually equivalent* relative to a story just in case assuming the information conveyed explicitly and implicitly by the story neither can be true if the other is false. Then the claim is that relative to their respective vignettes, (1) is contextually equivalent to (1′), (2) to (2′), and (3) to the negation of (3′). This is the sense in which our target sentences are tacitly modal – they are contextually equivalent to overtly modal ones.

Our hypothesis predicts, for example, that by and large people who read the first vignette thinking (1a) is true also think (1′a) is true, and that by and large people who read the third vignette thinking (3b) is false also think (3′b) is true. But the hypothesis does *not* make the stronger prediction that people will be inclined to judge that (1a) is true just in case (1′a) is, or that (3b) is true just in case (3′b) is false. In making judgments of equivalence people abstract from the contextual information they have, and so the semantic differences between the target sentences and their proxies come to the fore. Also, we are *not* saying that when people think about the truth or falsity of (1a) or (3b), they must consider the particular English sentences (1′a) and (3′b). That would amount to making the absurd prediction that people who know enough English to understand (1a) or (3b), but don't know what the words 'circumstances' or 'action' mean will react differently to the vignettes than the rest of us. What matters is that if the relevant contextual equivalence holds and we if can explain judgments about the proxies, those explanations carry over to the target sentences as well.

Let's sum up where we are. We have a two step plan to provide a unified and charitable explanation of the results of the experiments discussed in section 1. The first step is to reduce the problem of accounting for the impact of normative considerations on judgments about freedom, causation, and intentionality to the impact of normative considerations on modal judgment. The first step is now done: we have identified modal proxies for each of our target sentences. What remains is the second step: to explain why normative considerations impact the interpretation of the modal proxies.

## 3. Modal flavors

Our approach to modals will closely follow the classic framework introduced by Kratzer (1977). However, we will be arguing that this framework opens up a possibility that has not yet been adequately explored within the Kratzerian tradition. The claim will be that the framework actually allows for a simple explanation of the surprising ways in which normative considerations impact people's intuitions about modals.

But before we can begin sketching this new possibility, it will be necessary to briefly review Kratzer's original proposal.

### 3.1. From Ambiguity to Contextualism

Let us begin, then, with some well-known facts about natural language modals. One of the first things one notices about modal verbs like 'have (to)' is that they are used in slightly different ways in different sentences.[9] For example, in normal contexts 'have (to)' permits at least a circumstantial, a teleological, and a deontic reading:

(4) I had to sneeze during the talk.
(4′) Given my circumstances, I had to sneeze during the talk.

(5) You have to turn left at the next intersection.
(5′) Given your goals, you have to turn left at the next intersection.

[9] In English, as in other languages, modal expressions fall in a variety of syntactic categories. Following the literature, we are assuming that modal verbs ('have (to)', 'need (to)', 'ought (to)', 'dare (to)', etc.), modal auxiliaries ('must', 'can', 'might', 'should', etc.) and modal adverbs ('possibly', 'necessarily', 'probably', 'maybe', etc.) are all interpreted as expressions taking clausal scope. This assumption may well be incorrect but not in a way (we hope) that would matter for our purposes.

(6) You have to stop seeing that woman.
(6′) Given your moral obligations, you have to stop seeing that woman.

The standard, theory-neutral term for these different uses is *flavor*. One obvious approach to flavors would be to suggest that there is an ambiguity in 'have (to)': it has a circumstantial meaning a teleological meaning, a deontic meaning, and so on, for each of the different flavors we might uncover.[10]

Yet, though this approach may at first seem plausible, Kratzer shows that it cannot be the right one. Focus just on the alleged deontic meaning of 'have (to)' in (6). It seems that this sentence could still be used with subtly different meanings in different contexts. In one context, it might be a specifically moral appeal. But in another context, it could be a prudential one, paraphraseable as 'Given your best interests, you have to stop seeing that woman'. Or it could be an altruistic claim, as in 'Given her best interests, you have to stop seeing that woman'. Alternatively, the speaker may withhold judgment on what morality demands and on what is in anyone's best interest, voicing simply what company policy requires. At the bottom of this slope lies a theory postulating an indefinite number of deontic meanings for 'have (to)'. Let's not go there.

The obvious alternative is to say that all deontic modals share a single meaning and the differences one finds among their different uses are due to the influence of conversational context. The function of 'given'-clauses would be simply to make information relative to which they are to be understood more explicit.

But once we come this far it isn't easy to stop. Instead of positing separate meanings of 'have (to)' to explain the differences between different flavors of modality, Kratzer suggests that these differences, too, are to be explained in terms of context-dependence. For example, perhaps the only reason we think 'have (to)' is circumstantial in (4) and teleological in (5) is that we expect them to be used in different contexts. The expectation can be cancelled, in which case the interpretation shifts. In contexts appropriately primed, (4) will express a teleological necessity and (5) a circumstantial one:

[10] Besides these three, it is customary to talk about epistemic, bouletic, doxastic, and stereotypical flavors as well as of ability modals. Philosophers who believe in distinctive logical, conceptual, physical, or metaphysical modalities – and believe in addition that these are expressible in natural language – would add further items to the standard list.

[My friend three seats to my left fell asleep during the talk and I sneezed to wake him up and spare him of further embarrassment. Afterwards, I say:]

(4) I had to sneeze during the talk.

(4″) Given my goals, I had to sneeze during the talk.

[You are driving and I am holding a gun against your head. I say:]

(5) You have to turn left at the next intersection.

(5″) Given your circumstances, you have to turn left at the next intersection.

Kratzer's arguments here are highly persuasive, and it is now widely accepted that the differences between circumstantial, teleological and deontic modals should be understood not in terms of an ambiguity but rather in terms of a contextually given parameter. Kratzer herself takes this approach even farther – she suggests that the very same semantics to interpret epistemic modals as well:

(7) There has to be another copy of this book in the library.

(7′) Given what I know, there has to be another copy of this book in the library.

This last claim has been controversial. Some researchers have argued that while it might be possible to give a single unified theory of all of the non-epistemic modals (usually grouped together as *root modals*), epistemic modals are deeply different in important respects (Yalcin 2007; Gillies 2010). [11] We leave this controversy to one side and simply focus on root modals.

Our assumption, then, will be that when 'have (to)' is interpreted as a root modal the expression has a single meaning. The question now is what that meaning might be.

### 3.2. Modality and Quantificational Domains

To begin with, it may be helpful to introduce an analogy. Consider the quantifier 'all' and its use in sentences like 'All the beer is in the refrigerator.' Suppose this sentence is used by the host of a party when she sees one of her guests standing with an empty beer bottle looking around anxiously. Plausibly, in this context the sentence does not express the proposition that all of the beer in the entire universe is in the refrigerator. If we wanted a good paraphrase – a sentence that expresses more or less what the original sentence does without relying on the context as heavily as the original sentence does – we could say 'All the beer that is such-and-such (i.e. in the room,

[11] Kratzer (1991: 650) allows that syntactic differences between epistemic and root modals may correlate with differences in their argument structure.

around here, easily obtainable, designated for the party, etc.) is in the refrigerator'. The sentence quantifies over beer within a limited *domain* – not over all the beer there is, only over all the beer that is such-and-such.[12]

If modals are quantifiers over possibilities, we should expect that they too quantify over a contextually restricted domain. Those who judge that in the scenario described by our first vignette the sentence 'The captain had to throw the cargo overboard' is true do not believe that there are absolutely no possibilities of any kind in which the captain refuses to throw the cargo overboard; they presumably realize that such stubbornness would not against the laws of logic or even the laws of nature. A good paraphrase for the sentence would be 'In all possibilities that are such-and-such, the captain throws his wife overboard.' Here too, context has to select a domain for the sentence to quantify over – not a domain of all the possibilities that there are, only a domain of possibilities that are such-and-such.

There are good reasons to think that domain restriction for modals is more complex than domain restriction for quantificational determiners. Beer in a liquor store miles away is not in the domain because it is *irrelevant* in the context of helping out a party guest. But possibilities where the captain does not throw the cargo overboard are not at all irrelevant in the context of our first vignette. Such possibilities are directly invoked when it is said that 'the only way that the captain could keep the ship from capsizing was to throw his wife's expensive cargo overboard.' It would be quite implausible to assume that upon reading about what would happen if the captain failed to throw away the cargo people simply discard this possibility.[13] Intuitively, the necessity modal in 'The captain had to throw the cargo overboard' does not quantify over all the relevant possibilities, only over those that are in some sense *best* among the relevant ones. A slightly simplified version of Kratzer's semantics captures this insight as follows.

[12] This view is standard but not uncontested. For a defense, see Stanley and Szabó 2000. Many philosophers (e.g. Bach 1994) believe that 'All the beer is in the refrigerator' does not semantically express a proposition; some (e.g. Cappelen and Lepore 2005) contend that it expresses a minimal proposition whose truth-conditions we cannot spell out in non-disquotational fashion. These theorists agree that domains enter interpretation only at the level of ascertaining what the speaker uttering this sentence asserted in the context. If they are right, our story about domain restriction for modals should also be presented at that level. We do not believe that this would require substantive changes.

[13] An earlier version of this paper failed to take this insight into account and as a result operated within an oversimplified semantic framework. Thanks to Kai von Fintel for the criticism.

The sorts of things that can intuitively influence what possibilities a modal is quantifying over are the conversational participants' desires, goals, obligations or circumstances. Now, one's actual desires, goals, obligations and circumstances can all be represented by sets of propositions, and since one's desires, goals, obligations and circumstances could all be different from what they actually are, they should then be represented by functions from possibilities to sets of propositions.[14] This is exactly what Kratzer takes *conversational backgrounds* to be. Context determines two conversational backgrounds: one for settling which possibilities are relevant and another for settling which relevant possibilities are the best. Kratzer calls the former the *modal base* and the latter the *ordering source.*

Let $w$ be a possibility, $f$ a modal base and $g$ an ordering source. Then, we can say that a possibility $v$ is relevant according to $f(w)$ just in case all the propositions in $f(w)$ are true in $v$. (Say the modal base yields for $w$ the set of propositions stating Jack's obligations in $w$; then the relevant possibilities are those compatible with Jack fulfilling all his obligations in $w$.) We can also say that possibility $v$ is better than possibility $v'$ according to $g(w)$ iff there is a proposition in $g(w)$ that is true in $v$ but not in $v'$ and there is no proposition in $g(w)$ that is true in $v'$ but not in $v$. (Say the ordering source yields for $w$ the set of propositions that are Jill's goals in $w$; then a possibility is better than another iff there is a goal she has in $w$ that she realizes in the former but not the latter while all the goals she realizes in the latter she also realizes in the former.) Finally, we can say that $v$ is among the best relevant possibilities according to $f(w)$ and $g(w)$ iff there are no possibilities better than it according to $g(w)$ among the ones that are relevant according to $f(w)$. (Thus, if the modal base and the ordering source are as before, then $w$ is one of the best relevant possibilities iff (i) $w$ is compatible with Jack fulfilling all his obligations and (ii) there is no world $w'$ among the worlds compatible with Jack fulfilling all his obligations where Jill realizes all the goals she realizes in $w$ and some more.) It is worth mentioning that conversational backgrounds can also be empty (i.e. functions from possibilities whose range is the empty set). When an ordering source is empty, the best relevant possibilities are all the relevant possibilities; when the modal base is empty they are the possibilities that are best simpliciter.

[14] Ordinary quantifiers also range over different sets in different possible worlds. 'All the beer is in the fridge' can be true in the actual world, but false in a world where there is a perspicuous bottle of beer on top of the fridge, even if in the actual world that bottle happens to be is in the store, and hence outside the domain of quantification.

Assuming possibilities are worlds (an optional but standard assumption) and taking necessity modals to be operators (another optional but fairly common view), their semantic clause can be written as follows.[15,16]

(8) $[\![\textit{have to } \varphi]\!]^{w,f,g} = 1$ iff for all *v* among the best relevant worlds according to *f*(*w*) and *g*(*w*), $[\![\varphi]\!]^{v,f,g} = 1$

This sort of semantics can be seen as associating two domains with 'have (to)'. First, the modal base determines an *outer domain* – the set of worlds that are relevant for the interpretation of the modal. Second, the ordering source selects an *inner domain* – the subset of the outer domain that contains its best-ranked worlds. 'Have (to)' is a universal quantifier over the inner domain. [17]

'Given'-clauses function as explicit domain restrictors: possibilities incompatible with what is said to be given are all irrelevant in assessing a modal following the 'given'-clause. The simplest way to achieve this effect is to say that the semantic value of 'given α' (where α could be a noun phrase or a free relative) yields a set of propositions for any possible world. (You can think of expressions like 'given the storm' or 'given what the captain saw' as mapping a possible world to a set of propositions that are true in that world if the storm occurs in that world, or the set of propositions the captain saw in that world, respectively.) Then, the actual outer domain of a sentence of the form '*given* α, *have to* φ' consists of worlds where all propositions of the actual modal base as well as all the propositions within the semantic value of 'given α' are true. In other words, we have the following semantic clause:

(9) $[\![\textit{given } \alpha, \textit{have to } \varphi]\!]^{w,f,g} = 1$ iff for all *v* among the best relevant worlds according to $f(w) \cap [\![\textit{given } \alpha]\!]^{w,f,g}$ and *g*(*w*), $[\![\varphi]\!]^{v,f} = 1$

[15] For the record, we believe neither of these assumptions. We prefer to think of possibilities more along the lines of situations and we think modals are genuine quantifiers binding situation variables at the level of logical form. But these are some of the many semantic assumptions we do not wish to defend here – they have nothing to do with the topic of this paper.

[16] This assumes the limit assumption, i.e. that the set of worlds in the modal bases that are best-ranked by the ordering source is non-empty. Kratzer's actual semantics is more complex because she drops the limit assumption. Whether the more complex semantics ultimately captures our intuitions remains questionable (cf. Swanson 2010).

[17] One might wonder whether there is any reason to determine the inner domain by first extracting an order from a conversational background and then using the order to single out the best-ranked possibilities. The answer is yes – as Kratzer 1981 and 1991 argues, the order can be used to provide an elegant semantics for degree and comparative modals, such as 'there is a slight possibility that *p*' or 'it is more likely that *p* than that *q*'.

Thus, for example, 'Given the storm, the captain had to throw his wife overboard' is true in world *w* relative to modal base *f* and ordering source *g* just in case 'The captain throws his wife overboard' is true in every possible world that is best according to $g(w)$ among the worlds where all the propositions in $f(w)$ as well as all the propositions entailed by the occurrence of the storm are true.

## 4. Impurity

Thus far, we have been reviewing some of the basic elements of the classic Kratzer framework. We now want to suggest that this framework gives us just the resources we need to explain the surprisingly pervasive impact of normative considerations on people's intuitions about certain modals.

If modals actually were ambiguous, it would be natural to posit a list of possible flavors and to assume that any given modal had to fit neatly into one of them. Some modals would be purely circumstantial, some purely teleological, some purely deontic, but no single modal could include a mix of these different flavors. (It would make no sense, for example, to suppose that a given modal was best interpreted as being mostly teleological but also a little bit deontic.) We will refer to this view about the relationship between different flavors of modality as the assumption of *modal purity*.

As soon as one gives up the idea that modals are ambiguous and shifts instead to a theory based on context, the assumption of modal purity begins to look suspect. The most natural way of thinking about conversational contexts would be to assume that they embody a mixture of different information. In a given context, we might be primarily concerned with the circumstances but also somewhat concerned with the importance of achieving certain goals and very slightly concerned with the importance of not acting immorally. All these can impact which possibilities are relevant and thus, indirectly, which modal sentences are true. In other words, when one shifts from a theory based on ambiguity to a theory based on context, it is only natural to suppose that there can be *impure modals*.

Perhaps the best way of making sense of these phenomena is to take more seriously the metaphor of flavors. If one is asked to describe the flavor of a cake, one might well respond by saying

'chocolate.' But, of course, one does not thereby mean to suggest that the cake is composed entirely of chocolate and has no other flavors of any kind; the suggestion is rather that the dominant flavor of the cake is chocolate (though it may have other flavors as well). We suspect modal flavors are no different in this regard. In impure modals, the dominant flavor may be circumstantial or teleological, but there will always be at least a taste of the deontic there all the same.

Within the Kratzer framework, one could try to capture the impurity of modal flavors very conservatively. Let's say that a conversational background is pure just in case it assigns to each possible world a set of propositions that are, intuitively, of a single type. Conversational backgrounds associated with 'given'-clauses are typically like that. For example, 'given company policy' picks out in each possible world a set of propositions laid down in the company's policy book in that world and 'given what is expected from Bill' picks out in each possible world a set of propositions specifying the expectations towards Bill. Then we could try to maintain that although modals aren't always pure, their conversational bases are. We could say, for example, that 'ought' is associated with a purely circumstantial modal base and a purely deontic ordering source. Modal impurity would thus be the result of mixing two different pure considerations in determining the inner domain. Call this the assumption of *conversational purity*.

Conversational purity is a simple hypothesis, so it is worth spelling out why it cannot be right.[18] Consider first a case where we are driving to a party. You are at the wheel, we give directions. When you get to the next intersection, in principle you can do four things: drive straight, turn right, turn left, or turn back. In terms of getting to the party these options are ranked as listed: the best is to drive straight, the second best is to turn right, the third best is to turn left, and the worst is to turn back. Now suppose that the very best possibility is ruled out by the circumstances (you can't go straight because there is a fallen tree blocking the road), and the second best possibility

[18] As far as we can tell, there is nothing in Kratzer's writings that would commit her to conversational purity. Although the conversational backgrounds she explicitly defines are always pure, this might just be a matter of idealization. Recently, the assumption of conversational purity has come under scrutiny and a number of authors have suggested that the ordering source of deontic modals is "information-sensitive", i.e. that epistemic considerations may enter into the deontic ordering of circumstantially accessible possible worlds; cf. Kolodny and MacFarlane 2011, Charlow 2011, Cariani forthcoming. What we will propose to account for the experiments is complementary: we think normative considerations may enter into the non-deontic ordering of circumstantially accessible possible worlds.

is ruled out by the traffic laws (you can't turn right because there is a no right turn sign). We could then say 'You have to turn left at the next intersection.' This seems true. Turning left is *jointly* demanded by our goals, the circumstances, and the laws. We have three different kinds of considerations and only two conversational bases to accommodate them – so at least one of them must be impure.

The moral is simple: as long as we operate with a set of pure flavors – circumstantial, teleological, deontic, etc. – we cannot account for the semantics of modals in natural languages. Mixing two pure flavors is not enough: it is as though one recognized that there can be a chocolate cake with vanilla frosting, but then concluded that the cake itself was pure chocolate and the frosting pure vanilla. Modal impurity with conversational purity is still too pure to be true.

Abandoning conversational impurity, it is fairly simple to explain what is going on with our advice to turn left in the above example. The ordering source seems teleological: 'You have to turn left at the next intersection' tells you what is best in light of our goal of getting to the party. The modal base is mixed: it spells out the constraints we face in trying to achieve our goal. Some of these constrains are circumstantial (such as the fallen tree blocking the road ahead), others are normative (such as the rules of traffic banning a right turn). The dominant flavor of the sentence is teleological, but there is a normative hint. Possibilities where the car violates traffic rules are ignored because they are irrelevant: we take it for granted that we will not break those rules. This means that they end up outside the outer domain, which in turn makes them irrelevant when we assess what the worlds in the inner domain are like.

We now argue that a similar mechanism can explain each of the effects from the experimental literature reviewed above. The main difference is that unlike in the case just discussed, the impact of the normative is on the ordering source, not on the modal base.

## 5. The economy of hope

A purely deontic ordering source would rank possibilities according their conformity to a set of norms. But the ordering sources at play in the interpretation of our modal proxies are not pure: deontic considerations merely *constrain* the ranking they determine. Still, it may be possible to

articulate general principles that describe the way in which this happens. Our aim is to find such principles that are simple and are in line with ordinary judgments. We will introduce them as we go through our cases.

### 5.1. Freedom

Let's focus first on the effect for 'forced'. There, our example was (1), which we claimed had the modal proxy (1′).

(1) The captain was forced to throw the (a) cargo/(b) his wife overboard.
(1') Given the circumstances, the captain had to throw (a) the cargo/(b) his wife overboard.

On the account of modality discussed in the previous section, the proxy can be understood as quantifying over a certain set of possibilities, as made explicit in (1′′).

(1") Among the relevant possibilities compatible with the circumstances, in all the highest ranked ones the captain throws (a) the cargo/(b) his wife overboard.

To capture the intuitive judgments, we need to guarantee that the inner domain does not include possibilities in which the captain chooses not to throw the cargo overboard but does include possibilities in which the captain chooses not to throw his wife overboard. The former seems easy. It seems bizarre even to consider the possibility that the captain might choose not to throw the cargo overboard – given the circumstances, this seems like a rather far-fetched possibility. What we need is a principle that prevents us from ignoring the superficially analogous possibility where the captain chooses not to throw his wife overboard.

Note that the issue here is not whether the captain would ignore a possibility but rather whether we who are evaluating the sentence would do that. Even if the captain is a hardened wife-hating psychopath, the possibility of that he might not throw his wife overboard will still be deemed highly relevant by most people confronted with the scenario. What matters is not the probabilistic claim that he was in any way likely to do otherwise but the deontic claim that his actual behavior violated a moral rule.

Let's now say that a possibility is *hopeful* in a context just in case it is a possibility where none of the events at issue are salient norm-violations in the context. We propose the following principle:

*Hope*: The inner domain contains a hopeful possibility.[19]

*Hope* accounts for the contrast in (1″). Murder is a salient norm-violation even in the context of discussing what to do in a life-threatening storm, while destruction of property is not.[20] *Hope* ensures that in the context of the vignettes some possibilities where the captain refrains from throwing his wife overboard are in the inner domain and so (1″b) comes out as false. By contrast, *Hope* makes no predictions about (1″a). It is judged true presumably because possibilities where the captain refuses to toss the cargo overboard are deemed sufficiently far-fetched in the circumstances. They are excluded from the outer domain, and consequently are not in the inner domain either.

You can think of *Hope* as an implementation of an *ought implies can* principle.[21] Salient norm-violations ought not to occur and – if the inner domain is bound to contain a possibility where they don't – salient norm violations can fail to occur. Accordingly, the principle runs into difficulties when it comes to genuine dilemmas. Suppose the captain has a choice: he can save the boat by throwing any one of the passengers overboard. Is he forced to do so? As long as *Hope* is in effect, we predict that people will say no. Perhaps they imagine that there is a way to avoid the killing and still save the ship even if this is ruled out explicitly. But there are limits to this – if there is no way to dismiss the problem and no way to make the hopeful possibility relevant many will abandon *Hope* and resolve the dilemma one way or another. How people *actually* do that is an important question for psychology, how they *should* do that is an important question for ethics. We take no stance on either. What we claim is that people do not normally abandon *Hope* and this fact can explain the results in this experiment.

---

[19] *Hope* establishes a connection among different features determined by the same context – the events at issue that are salient norm-violations, the modal base, and the ordering source. In this regard, it is similar to the principle that says that the speaker of the context is located at the place and time of the context. This principle is responsible for 'I am here now' coming out as true in any context. Similarly, *Hope* ensures that 'This did not have to happen' comes out as true in any context where 'this' refers to a contextually salient norm-violation. Just as there are cases when 'I am not here right now' comes out intuitively true (think of answering machines) there are also cases when it appears we can truthfully say 'This crime had to happen' (think of defense attorneys). Such cases may be handled by allowing special contextual features to override the relevant principles at the level of what is said, or perhaps at the level of what is communicated.

[20] Note that the claim is not that the captain does not violate a salient norm when he throws the cargo overboard. He surely does, which is why it is proper for him to deliberate before doing so. But given the circumstances, what he chooses to is not a salient norm-violation.

[21] We thank Tad Brennan for this observation.

The status of *Hope* is similar to the rules Lewis (1996) introduces in his work on knowledge ascriptions (*Rule of Actuality*, *Rule of Attention*, *Rule of Reliability*, etc.). Like those, *Hope* is a principle that governs the domain of possibilities we quantify over when we make overt or covert modal claims. Lewis's rules turned out to be making some pretty bad predictions, so they tend to be rejected today even by those who are largely sympathetic towards his general outlook.[22] This may well happen in time to *Hope* as well, which would be unfortunate but not terribly so. What matters for us is contextualism about modality (which is widely accepted in semantics) and impurity (which is not as widely accepted). These are what make it possible to explain the data of section 1 by appealing to some domain principle or other.

### 5.2. Causation

Let's turn now to the effect for 'cause'. There, our example was (2), which we claim has the modal proxy (2′), which in turn has the quantificational paraphrase (2′′).

(2) (a) The professor/(b) The assistant caused the problem.
(2') Given the action of (a) the professor/(b) the assistant, the problem had to arise.
(2′′) Among the relevant possibilities compatible with (a) the professor's/(b) the assistant's action, in all the highest ranked ones the problem arises.

The experimental data show that people agree with the claim that the professor caused the problem (2a) but disagree with the claim that the assistant caused the problem (2b). The task now is to explain this asymmetry in terms of the inner domain generated by the modal base and the ordering source. More specifically, we need to show that this domain contains no possibility where the professor takes a pen and the problem fails to arise but contains a possibility where the assistant takes a pen and the problem fails to arise.

The latter seems easy: surely it could have happened that the professor refrains from taking a pen, in which case whether or not the assistant takes one, there would be no problem at the desk. But the former is a puzzle: even if the professor does take a pen, the assistant might still refrain from taking one, in which case, again, the problem would not arise. *Hope* alone is clearly no help – it guarantees that the inner domain includes certain worlds but what we need here is a

[22] Section 6 of Stanley 2005 discusses some particularly acute problems with the *Rule of Belief* and *the Rule of Actuality*.

guarantee that it fails to contain certain worlds. To explain the asymmetry here, we will need to introduce further resources.

Our strategy will be to make use of an approach that has proven helpful in numerous other areas: an appeal to economy. The economy principle is motivated by the fact that in assessing modal claims the domain has to be surveyed, which takes genuine cognitive effort:

*Economy*: The outer domain is the smallest one that satisfies all other principles.

The assumption here is that there are a variety of principles governing relevance. Some of these say that certain possibilities are irrelevant – e.g. other things being equal far-fetched options are not in the domain. Others do the opposite, ensuring certain possibilities are not ignored – e.g. other things being equal, possibilities that are explicitly mentioned are within the domain. These are first-order principles because they tell us whether possibilities of a certain type are relevant or not. *Hope* is one of the first-order domain principles. But first-order principles by themselves cannot fix the domain. They can and often do come into conflict with one another, and when they don't, they typically severely underdetermine what relevant possibilities there are. So, we need some meta-principles in addition. *Economy* is one of these; it says that we should select the conversational backgrounds that determine the smallest domain that satisfies all other principles.

It should now be possible to see, at least in broad outline, how one might explain the judgments in the pen vignette. The action of the professor is a salient norm-violation, so by *Hope*, the domain must include at least one possibility in which it does not occur. For this reason, we include in the inner domain a possibility in which the professor does not take a pen. Given all we are told about the situation in the vignette, we know this must be a world where the receptionist does not run out of pens, so (2″b) comes out false. However, there is no principle mandating that we include in the domain a possibility in which the administrative assistant does not take a pen. Thus, by *Economy*, some of these possibilities are eliminated from the domain, which leads to the prediction that (2″a) is true.

Let's see how this works in more detail. We have three binary choices – whether the professor takes a pen, whether the assistant takes a pen, and whether the problem at the desk arises. These generate eight types of possibilities:

| | *w1* | *w2* | *w3* | *w4* | *w5* | *w6* | *w7* | *w8* |
|---|---|---|---|---|---|---|---|---|
| Professor takes pen | + | + | + | + | – | – | – | – |
| Assistant takes pen | + | + | – | – | + | + | – | – |
| Problem at the desk | + | – | + | – | + | – | + | – |

While the vignette does not *say* this, it *strongly suggests* that four of these eight possibilities (*w2*, *w3*, *w5*, and *w7*) are irrelevant, i.e. not included in the outer domain. This is so because if any of these worlds is actual, the vignette is intuitively incomplete. Take *w2*. It is indeed possible for both the professor and the assistant to take a pen and the receptionist still having one left to take a note. Perhaps she has a secret stash of pens in her drawer which she regularly relies on in cases of emergency. Alas, on the Monday morning described she inexplicably found her drawer empty, which is why she was unable to take the message. Of course, if all this is true it is decidedly odd that the vignette is silent about crucial details. Assuming that the vignette is not misleading, there cannot be no secret stash and we can rule out *w2* as a relevant option. In *w3*, the secretary runs out of pens even though the assistant declines to take one. What happens to the pen the assistant took in the world the vignette describes? Perhaps it ran out of ink and the receptionist discarded it before she received the phone call. But then we should wonder why the vignette fails to mention this fact. So again, assuming the vignette is not unduly reticent, *w3* must be an irrelevant possibility. Similar broadly Gricean considerations rule out *w5* and *w7* as well.

Of the remaining possibilities *w1* must be surely be included in the inner domain on the grounds that it is the one that according to the vignette occurs. *w6* and *w8* are the only remaining ones that are hopeful, i.e. where the professor does not take a pen. According to *Hope*, at least one of them must be included in the inner domain and since there appears to be no principled basis to select one over the other, presumably both are in. Finally, *w4* is not hopeful and – assuming there is no further principle that requires that we take it into account – by *Economy* it is irrelevant; i.e. it is not in the outer domain. Thus, the domain consists of *w1*, *w6* and *w8*, which means that (2a″) comes out true and (2b″) false.

It is worth noting that the explanation we provided is completely neutral with respect to the question as to whether the outcome itself is a salient norm-violation. In this particular case, the

outcome is something bad (the receptionist having a problem), but the explanation would go through in exactly the same way even if the outcome had not been bad at all. We can still exclude *w2*, *w3*, *w5*, and *w7* from the outer domain on Gricean grounds; we must still include *w1* in the inner domain because it is actual; we get still have only *w6* and *w8* as the only hopeful possibilities; and we can still eliminate *w4* an economy grounds.

This neutrality is an important virtue of the explanation. Suppose we modified our case in such a way that the outcome ended up being something good (e.g., it turned out to be extremely helpful that there were no pens on the desk). The theory now generates the seemingly paradoxical prediction that the person who acted wrongly will specifically be singled out as the cause of the good outcome. In fact, that is precisely the result obtained in experimental studies using cases of this form (Hitchcock and Knobe 2009).

As our example illustrates, *Hope* and *Economy* are not the only principles governing domain selection. We relied on the substantive principle that a possibility known to be actual must be relevant and on the meta-principle which forbids making arbitrary distinctions among possibilities. Presumably, a variety of other principles are also at work in domain selection.[23]

[23] Consider, for example, the sentence: 'Given the action of the professor, the administrative assistant's action had to occur.' Intuitively, this would be false in the context of the vignette describing the events of the missing pens. But if the inner domain does not include *w4*, we predict that the sentence is true. (Thanks to Brian Weatherson and Stewart Shapiro for the objection.) There must be some principle that prevents the elimination of this possibility when this sentence is evaluated. Actually, there are probably two. The first is another economy principle which says that that you should keep the events at issue at a minimum. Thus, when you are evaluating 'Given the action of the professor, the action of the assistant had to occur' you can forget about the problem at the desk. We have four kinds of possibilities to consider: (i) both takes pens, (ii) only the professor does, (iii) only the assistant does, and (iv) neither does. (i) is actual, so it is in the inner domain. By *Hope*, either (iii) or (iv) are in the inner domain, and since there is no basis to choose, both are. One might think that *Economy* lets us cut (ii) from the outer domain, but this is where the second principle comes to play. This is a non-triviality principle which says that no modal evaluation can depend exclusively on economy considerations (e.g. you cannot judge a sentence to be necessary simply by ignoring all the possibilities of it being false). Now, if we dropped (ii), 'Given the action of the professor, the action of the assistant had to occur' would turn out to be true simply because we are ignoring all the possibilities that could falsify it. So, (ii) cannot be eliminated and the sentence is false. Note that we did not violate this non-triviality principle in evaluating 'Given the action of the professor, the problem at the desk had to occur'. That's because there we had three events at issue generating eight possibilities, of which two – *w3* and *w4* – can falsify the sentence. We did exclude *w4* on the basis of *Economy*, but *w3* – the possibility where the professor takes a pen, the assistant does not and the problem still arises – was excluded on Gricean grounds.

### 5.3. Intentionality

The effect for attributions of intentionality was illustrated with (3), which has the negative modal proxy (3′) paraphraseable as (3′′):

(3) Jake hit (a) the bull's eye/(b) his aunt intentionally.
(3') Given the fluke, Jake had to hit (a) the bull's eye/(b) his aunt.
(3′′) Among the relevant possibilities compatible with the fluke, Jake hit (a) the bull's eye/(b) his aunt, in all the highest ranked ones.

What needs explaining is why the inner domain contains no possibility where the fluke occurs and Jake still fails to hit the bull's eye (making (3a) false) but does contain a possibility where the fluke occurs but Jake still fails to hit his aunt (making (3b) true).

Our explanation is that the moral difference between the cases leads to a difference in which possibilities end up in the inner domain. In the bull's eye case, the only possibilities in the inner domain in which Jake does not hit the bull's eye are cases in which the fluke happens not to occur. For this reason, it appears that Jake's act of hitting the bull's eye is due entirely to luck, and it is not regarded as intentional.

In the aunt case, however, a different possibility becomes relevant. Since it was morally wrong of Jake to aim at his aunt, it becomes relevant to consider the possibilities in which Jake does not aim at his aunt in the first place. In this case, then, it appears that his act is due not only to luck but to his decision, and the act is therefore regarded as intentional.

Let us now spell this argument out in more detail for the case of the aunt. Think of the shooting scenario as involving three events: Jake aiming at his aunt, the fluke, and Jake's aunt being hit. The first of these is a salient norm-violation, the second is not; their joint occurrence is sufficient for the third. Here is the possibility chart:

| | *w1* | *w2* | *w3* | *w4* | *w5* | *w6* | *w7* | *w8* |
|---|---|---|---|---|---|---|---|---|
| Jake aims at his aunt | + | + | + | + | – | – | – | – |
| Jake's hand slips, gun stays on target | + | + | – | – | + | + | – | – |
| Jake's aunt is hit | + | – | + | – | + | – | + | – |

*w1* is in the inner domain because it is actual. *w2* is a world where Jake's aunt is miraculously escapes being hit despite being shot at from a gun that is on target; *w5* and *w7* are worlds where she is hit even though Jake doesn't even aim at her. These are certainly far-fetched possibilities, so they are outside of the outer domain. *w3* is a world where Jake aims at his aunt, the fluke does not occur, and he hits her. *w4* is similar, except that Jake misses his aunt. Unless there are independent reasons to consider them, *w3* and *w4* are excluded from the outer domain on grounds of economy. Of course, in this case there may well be independent reasons to consider *w3* and *w4*: the fluke is an extremely unlikely event, so perhaps possibilities where it does not occur cannot be ignored willy-nilly.

Be that as it may, the key point is that *w6* and *w8* are in the inner domain. Jake's actual behavior violates a salient norm, and *Hope* says that we need to include in the inner domain at least one possibility in which this norm is not violated. Both *w6* and *w8* are hopeful, and choosing between them would be arbitrary. We thereby arrive at an inner domain that includes possibilities in which the fluke does occur but Jake does not hit his aunt. For this reason, we predict that (3″b) is false.

**5.4. Remarks on Explanatory Strategy**

Perhaps it will be helpful here to pause for a moment and reflect on the general explanatory strategy we have been pursuing. In characterizing the modals under discussion here, one might have expected to find a simple and exceptionless rule that would easily handle all cases. The central idea of impurity is that no such rule exists. Accordingly, we have adopted a somewhat different explanatory strategy. We have posited a heterogeneous set of principles that, together, purport to account for people's intuitions in these cases. This explanatory strategy is then, by its very nature, an open-ended one. Although we have described certain principles here, it should be clear that there are numerous others still to be described.

At this point, one might well complain that our explanatory strategy gives us too much wiggle room. (Whenever a potential counterexample comes up, we can always wiggle out of it by positing a new principle or by manipulating the set of initial possibilities.) This complaint is in one way accurate and in another completely misguided. It is true that our explanatory strategy

allows us to escape a certain kind of burden. Since we claim that conversational backgrounds are impure and cannot be captured by a simple rule, we do not take on the burden of giving a single rule that will capture the data in all cases. But at the same time, we take on another burden that earlier accounts have shirked. When we are trying to explain the data about, say, 'cause,' we cannot introduce ad hoc principles that apply just to this one expression. Instead, we are forced to explain all of the data in terms of general principles of domain restriction that will have testable implications for numerous different expressions.

## 6. Conclusion

The aim of this paper was to present an explanation for the impact of normative considerations on people's assessment of certain seemingly purely descriptive matters. A number of experiments in the last few years have shown that people's judgments about whether an action was free or forced, whether it caused a certain outcome, and whether it was performed intentionally often depend on whether the action violates a norm. The explanation we provided is unified and charitable: we argued that there is a common core of the phenomenon and that these judgments are not in error.

The explanation is based on two main claims. First, a large category of expressions of prime philosophical concern are tacitly modal: they are contextually equivalent to modal proxies. Second, natural language modals can have impure flavors, being evaluated against conversational backgrounds shaped by heterogeneous considerations, including normative ones. This impurity in the conversational backgrounds can be modeled using a pair of principles which jointly make certain hopeful possibilities relevant at the expense of less hopeful ones.

Although we introduced the notion of impure modals as a way of explaining intuitions about freedom, causation and intentionality, modal impurity is also a phenomenon of interest in its own right. Future research might directly examine impure modals, quite independently of the ways in which such modals might be related to intuitions about other expressions of interest. One important question here is whether all natural language modals are impure or whether natural language also includes modals of pure flavors. A related question concerns the lexical restrictions on certain modal expressions (e.g., the way in which 'should' can be used as a

deontic modal but not as a circumstantial modal). One wants to know whether certain kinds of lexical restrictions can lead to modals with pure flavors.

Regardless of how these issues are resolved, the present study seems to be suggesting something surprising about the relationship between people's judgments about how things are and their judgments about how things ought to be. Hume famously claimed that it is "altogether inconceivable" that a proposition where the subject is connected to the predicate by an *ought* or an *ought not* could be derived from propositions where the connections are made by an *is* or an *is not* (*Treatise* 3.1.1.). While many philosophers would dispute that the chasm is that deep, it is received view that normative and descriptive considerations usually are, and always should be sharply distinguished from one another. If morality impacts our sound judgment about matters of freedom, causation, and intentionality, the received view is called into question. The challenge is not whether we can coherently draw the line between the normative and the descriptive. It is, rather, whether the distinction we know and cherish is as deeply rooted in ordinary thinking as it is often assumed. If our explanation of the phenomena is on the right track the answer to this question appears to be negative.

**References**


Adams, F. and A. Steadman 2004. Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding? *Analysis,* 64, 173-181.

Alicke, M. 2008. Blaming Badly. *Journal of Cognition and Culture*, 8, 179-186.

Bach, K. 1994. Conversational Impliciture. *Mind and Language* 9.124 – 62.

Cappelen, H. and E. Lepore 2005. *Insensitive Semantics.* Oxford: Blackwell.

Cushman, F., J. Knobe and W. Sinnott-Armstrong 2008. Moral appraisals affect doing/allowing judgments. *Cognition* 108:353-80.

Dretske, F. 1970. Epistemic Operators. *Journal of Philosophy* 67: 1007-23.

Egan, A., J. Hawthorne, and B. Weatherson 2005. Epistemic Modals in Context. In G. Preyer and G. Peter eds., *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. 131 – 68. Oxford: Oxford University Press.

Gillies, T. 2010. Iffiness. *Semantics and Pragmatics* 3: 1 – 42.

Guglielmo, S. and B.F. Malle 2010. Enough Skill to Kill: Intentionality Judgments and the Moral Valence of Action. *Cognition* 117: 139 – 150.

Guglielmo, S. and B.F. Malle 2011. The Timing of Blame and Intentionality: Testing the Moral Bias Hypothesis. Unpublished manuscript, Brown University.

Halpern, J. and J. Pearl 2005 Causes and Explanations: A Structural-model Approach — Part I: Causes. *British Journal for the Philosophy of Science* 56: 843–887.

Hintikka, J. 1962. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca, NY: Cornell University Press.

Hitchcock, C. 2007. Prevention, Preemption, and the Principle of Sufficient Reason *Philosophical Review* 116: 495 – 532.

Hitchcock, C. and J. Knobe 2009. Cause and Norm. *Journal of Philosophy* 11: 587 – 612.

Knobe, J. (2003). Intentional Action in Folk Psychology: An Experimental Investigation. *Philosophical Psychology,* 16, 309-324.

Knobe, J. 2010. Person as Scientist, Person as Moralist. *Behavioral and Brain Sciences* 33: 315 – 29.

Knobe, J. and Fraser, B. 2008. Causal Judgment and Moral Judgment: Two Experiments. In W. Sinnott-Armstrong ed., *Moral Psychology Vol. 2: The Cognitive Science of Morality: Intuition and Diversity*. Cambridge, MA: MIT Press.

Kratzer, A. 1977. What 'Must' and 'Can' Must and Can Mean. *Linguistics and Philosophy* 1: 337 – 55.

Kratzer, A. 1981. The Notional Category of Modality. In Eikmeyer, H.-J. And H. Rieser eds., *Words, Worlds, and Contexts.* 38 – 74. Berlin: de Gruyter.

Kratzer, A. 1991. Modality. In von Stechow, A. and D. Wunderlich eds., *Semantics: An International Handbook of Contemporary Research.* 639 – 50. Berlin: de Gruyter.

Lewis, D. 1973. Causation. *Journal of Philosophy* 70: 556 – 67.

Lewis, D. 2000. Causation as Influence. *Journal of Philosophy* 97: 182 – 97.

Lewis, D. 1996. Elusive Knowledge. *Australasian Journal of Philosophy* 74: 549 – 67.

Nadelhoffer, T. 2004. The Butler Problem Revisited. *Analysis* 64: 277 – 284.

Nadelhoffer, T. 2005. Skill, Luck, and Action. *Philosophical Psychology* 18: 343 – 54.

Nadelhoffer, T. 2006. Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality. *Philosophical Explorations* 9: 203 – 19.

Nanay, B. 2010. Morality or Modality? What Does the Attribution of Intentionality Depend On? *Canadian Journal of Philosophy*. 40: 25 – 39.

Nichols, S. and J. Ulatowski 2007. Intuitions and Individual Differences: The Knobe Effect Revisited. *Mind and Language*, 22: 346 – 65.

Phillips, J. and J. Knobe 2009. Moral Judgments and Intuitions about Freedom. *Psychological Inquiry*, 20, 30 – 6.

Portner, P. 2009. *Modality.* Oxford: Oxford University Press.

Roxborough, C. and J. Cumby 2009. Folk Psychological Concepts: Causation. *Philosophical Psychology* 22: 205 – 13.

Sousa, P. and C. Holbrook 2010. Folk Concepts of Intentional Action in the Contexts of Amoral and Immoral Luck. *Review of Philosophy and Psychology*.

Swanson, E. 2010. On the Treatment of Incomparability in Ordering Semantics and Premise Semantics. *Journal of Philosophical Logic.* Online First DOI: 10.1007/s10992-010-9157-z.

Sripada, C. and S. Konrath forthcoming. Telling More than We Can Know about Intentional Action. *Mind & Language*.

Stanley, J. 2005. *Knowledge and Practical Interests*. Oxford University Press.

Stanley, J. and Z. G. Szabó 2000. On Quantifier Domain Restriction. *Mind and Language* 15: 219 – 61.

Stephenson, T. 2007. Judge Dependence, Epistemic Modals, and Predicates of Personal Taste. *Linguistics and Philosophy* 30: 487 – 525.

Sytsma, J., J. Livengood and D. Rose 2010. Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions.

Szabó, Z.G. 2006. Sensitivity training. *Mind and Language* 21: 31 – 8.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Yalcin, S. 2007. Epistemic Modals. *Mind* 116: 983 – 1026.

Yalcin, S. forthcoming. Nonfactualism about Epistemic Modality. In Egan, A. and B. Weatherson eds. *Epistemic Modality.* Oxford University Press.

Young, L., F. Cushman, R. Adolphs, D. Tranel and M. Hauser 2006. Does Emotion Mediate the Effect of an Action's Moral Status on its Intentional Status? Neuropsychological Evidence. *Journal of Cognition and Culture* 6: 291 – 304.

Zalla, T., E. Machery and M. Leboyer forthcoming. Intentional Action and Moral Judgment in Asperger Syndrome and High-Functioning Autism. *Review of Philosophy and Psychology.*