To appear in *Philosophical Studies*

A Simple Linguistic Approach to the Knobe Effect, or the Knobe Effect without any Vignette

0. Introduction

In this paper I will propose a simple linguistic approach to the Knobe effect, or the moral asymmetry of intention-attribution in general.[1] Joshua Knobe has taken such an effect to be a psychological phenomenon, over and above what can be investigated through the traditional conceptual analysis (e.g. Knobe 2010, Phillips, Luguri, Knobe 2015). If he is right, it seems that language is not really relevant to the phenomenon and the linguistic approach to it is misguided from the start.[2] So let us first briefly discuss this issue starting from the linguistic relevance of experimental philosophy in general.

1. What are experimental philosophers doing?

Since most experiments in the experimental philosophy literature consist of the questionnaire survey, which of course uses and relies on language, it is natural to expect that they are investigating people's linguistic intuitions. If so, experimental philosophy, at least of its *positive program*, is empirically supplementing the traditional project of conceptual analysis. For example, articulating the positive and negative programs of experimental philosophy, Stich and Tobia (2016) say that experimental philosophers in

[1] Note that, even though the Knobe effect has often been also called the *side-effect effect*, the moral asymmetry of intention-attribution has been observed even when the vignette does not involve side-effects (e.g. Knobe 2003b, Nadelhoffer2005).

[2] Note that Strickland, Fisher, Knobe, and Keil (2015) reports that *syntax* affects our intentionality judgements. Thus, in this sense language is relevant to the intention-attribution. But we are here concerned with the relevance of the semantic aspect or concepts captured by the relevant expressions.

the positive program "are motivated to explore intuitions experimentally because they think that by doing so they can do a better job of conceptual analysis" (p. 33).[3] Let us call this the *supplementary picture* of experimental philosophy.[4]

However, although there Stich and Tobia use the Knobe effect as one of the representative results in the positive program of experimental philosophy, Knobe himself began to explicitly reject this kind of picture after his 2003a,[5] saying "we abandon the assumption that the study of people's intuitions about cases can only have philosophical significance insofar as it helps us to answer semantic questions" (Knobe 2007, p. 120). He instead claims that his research is about folk psychology, which is supposed to reveal facts about (in his words) "how the mind works" (*ibid.*).[6]

The main reason for not adopting the semantic/conceptual interpretation of the Knobe effect, or the moral asymmetry of intention-attribution, may be that it is part of

---

[3] On the other hand, experimental philosophy of the negative program denies the reliance on intuitions in philosophy in general, or in its weaker form, the reliance on intuitions without any empirical support. Weinberg (2016) takes the latter as *the* negative program of experimental philosophy. If so, it corresponds to Experimental Restrictionism of Nadelhoffer and Nahmias (2007). See section 5.2 of Weinberg (2016) for more on the negative program.

[4] Nadelhoffer and Nahmias (2007) called such a project *Experimental Analysis*, and Machery (2016, sec. 33.2) called it (in philosophy of science) *experimental conceptual analysis*. But note that, the supplementary picture can also be shared by armchair philosophers, and moreover, *pace* Knobe (see below), whether a paper or the result reported there is a contribution to conceptual analysis or not is not an inherent property of the paper or the result, but is rather a matter of how we use such result. Thus, K. Mortensen and J. Nagel use various results concerning, e.g., free will (Nahmias et al. 2006, Nichols and Knobe 2007, Murray and Nahmias 2014), phenomenal state ascription (Buckwalter and Phelan 2014), and the pain paradox (Reuter, Phillips, and Sytsma 2014), to show that "empirical work can extend and enhance the reach of traditional philosophical theorizing" (Mortensen and Nagel 2016, p. 64). In general, what Mortensen and Nagel say in section 4.5 of their 2016 (especially about what is called the "neutral" project there) is closely related to what I argue in this section.

[5] Though in his 2003a, Knobe heavily emphasizes implications of his findings for the *concept of intentional action* as their significance.

[6] See also sec. 2.3 of Knobe and Nichols (2008), and p.2 of Alexander (2012). This is called "Experimental Descriptivism" by Nadelhoffer and Nahmias (2007), and the "neutral" project by Sytsma and Machery (2013).

very general phenomena (Knobe 2010): Moral asymmetry has been observed not only in intention-attribution, but also attribution of mental states in general (deciding, desiring, being in favor of, advocating, knowing, believing, etc.), and even judgments on the doing/allowing distinction, causation, and freedom.[7] Recently, Knobe and his colleagues (hereafter PLK) provide a unifying account of all these phenomena in terms of the underlying general cognitive processes, rather than consulting individual concepts (PLK 2015). Given the scientific rigorousness and the large-scale of the studies reported there, one might even think that their conclusion is decisive and almost everything has been done on the topic.

More recently, Knobe (2016) again makes a general claim that the vast majority of the work of experimental philosophy does not even *intend* to contribute to conceptual analysis, claiming that only 10.4% of the studies in the PhilPapers database from 2009 to 2013 can be counted as defending a specific conceptual analysis.[8] Certainly, as many armchair philosophers would acknowledge, the results of experimental philosophy have contributed to the understanding of various philosophical concepts. But even so, that may be just a *side effect*, and experimental philosophers do not do it *intentionally*, according to Knobe.[9] For him, most of experimental philosophers are, as cognitive scientists, only interested in various psychological effects themselves, or the cognitive processes behind them.

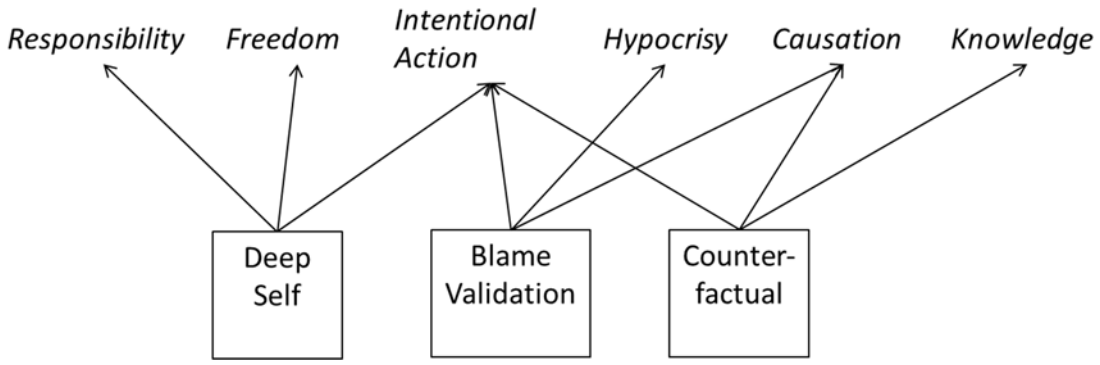
If Knobe is right about this, however, does he also mean to *discourage* the project of conceptual analysis in general as useless? If he does, he at least provides no argument

---

[7] For references of such literature, see section 1 of Robinson, Stey, Alfano (2015), and PLK (2015).
[8] But note the caveat concerning this categorization in footnote 4 above.
[9] So, if Knobe thinks that contributing to conceptual analysis is a morally good thing, he himself might have been affected by the Knobe effect.

against the supplementary picture. He certainly thinks that conceptual analysis is unproductive, and, to show this, appeals to the historical fact that the results of conceptual analysis (or the resultant theories based on it) have been messy, often leading to "monstrous complexity", without theoretical virtues of simplicity and elegance. Even if this was a fair verdict,[10] however, that does not prohibit him from claiming that revealing the underlying cognitive processes is a *better* approach to understand or investigate the relevant concept.

Indeed, the apparent assumptions behind Knobe's argument against the supplementary picture are questionable. He presents the following figure, where a series of concepts are on the top row and a series of underlying cognitive processes are on the bottom row.



<figure 1 (from Knobe 2016, Figure 3.1 of sec. 3.4)>

[10] It seems that, especially if there are *also* influences from concepts to cognitive processes (see below), the virtue of the simplicity of the theory can no more be expected for the theory of cognitive process than the one based on conceptual analysis. No doubt Knobe proposes an independent criterion of simplicity of explanation of a particular cognitive process relative to a complex general theory (*ibid*. sec. VII). But, then, the same criterion can and should be applied to conceptual analysis. Note, however, it is in fact acceptable for us to admit that no simple theory can be expected in either area. In such a case, Wittgensteinian quietism may prevail, and all the philosophical problems are to be *dissolved*, rather than solved. This will be done by understanding better and more perspicuously our own form of life. Indeed, after we obtaining the results about the particular cognitive processes, we have the right, or even the duty, to ask *why* we have such (rather than other) cognitive processes, which should be answered by our social facts lying outside of the head of individuals.

This figure seems to assume that: 1) the clear distinction between concepts and the underlying cognitive processes,[11] and 2) the unilateral influence (represented by the arrows) from the cognitive processes to the intuitions of relevant concepts.[12] It is unlikely that Knobe is really *committed to* such theses. However, once he explicitly accepts the denials of them, especially the interaction between concepts and cognitive processes, he should also admit the contribution of the study of concepts to the theory of cognitive process. Indeed, the study of relevant concepts is arguably a necessary condition for the satisfactory theory of a particular cognitive process, since we need to delineate to what extent the relevant concepts are, and from which point the cognitive process is, responsible for the data. And importantly, this line may vary from language to language (as the argument from linguistic diversity would suggest. See below).

Thus even if experimental philosophers are not subjectively committed to the investigation of concepts, such studies (empirical or *a priori*) are by no means incompatible with what they are doing as a matter of fact, and they should even contribute to each other, or so I shall argue below, thereby defending a broadly supplementary picture of experimental philosophy. In particular, I will argue (with empirical data) that

---

[11] Assumption 1 invites a variant of the notorious problem of drawing a clear line between meaning and belief, discussed by Donald Davidson (for his claim of the interdependence of belief and meaning, see Davidson 1973, p. 134, 1975, p. 158). If Davidson is right, concepts and cognitive processes should be likewise intertwined. This is also a matter of whether the relevant norms (governing the use of the relevant term) are at the linguistic level or psychological level. But in either case, at least for the externalist about meaning and content they exist outside of the head of individuals as social conventions.

[12] Note that the mutual influence relation has been exactly what is implied by the moral asymmetry in question. Intention-attributions, causal judgments, and others do have direct consequences on our moral judgments. What Knobe and others have revealed was that there is *also* the influence of the opposite direction.

there is still a room for a systematic linguistic investigation of the phenomena like the Knobe effect.

One way to reveal the linguistic relevance of the questionnaire research of experimental philosophy is to show the *linguistic diversity* of the relevant concepts. If, for example, we found that there are various concepts of knowledge captured by knowledge verbs in different languages, then Anglophone epistemologists cannot naïvely assume the concept captured by *English* "know" to be *the* concept of knowledge, and should at least be aware of its linguistic contingency.[13] Let us call this *argument from linguistic diversity*. However, even if we found the difference in answers between speakers of two languages in the survey using, say, Knobe's chairman case, the question remains: Is the difference to be explained by the linguistic difference, or the cultural-psychological one? This is in fact the same problem as the one Knobe's assumption 1 does (if he really assumes it).[14]

Still, we can at least try to keep the influence of psychological factor minimal. One way to do so is to eliminate the whole vignette Knobe provided, and ask merely linguistic judgments of ordinary people, about the felicity of the use of "intentionally" in various sentences, in particular, of course, sentences like "X intentionally harmed the environment" and "X intentionally helped the environment". If we find the asymmetry analogous to the Knobe effect in judgments on these sentences, then the Knobe effect is at least partly due to the use of linguistic expression "intentionally", or the concept of *intention* behind it. If, however, the same method reveals the linguistic diversity of the

---

[13] For example, Japanese has two knowledge verbs "shitte-iru" and "wakatte-iru", and it has been reported that they express quite different concepts even when used for propositional knowledge (Mizumoto *forthcoming*). If so, this should support pluralism of the concept of knowledge.

[14] Note also that, Davidson's argument for the interdependence of meaning and belief was presented in the *cross-linguistic* context, or that of radical interpretation.

concept of intention, that will further bolster the claim of the linguistic relevance of the Knobe effect and other moral asymmetries of intention-attribution.

2. Felicity judgment surveys

Thus, rather than asking for judgments about the action of the protagonist in a particular situation by providing some vignette, we collected people's linguistic judgments on bare sentences, through asking judgments about the felicity of them. In view of the argument from linguistic diversity, for additional evidence we planned the same survey with Japanese subjects using Japanese. So, even if we found no asymmetry in judgements of English speakers on the relevant sentences, if we found the asymmetry (analogous to the Knobe effect) in the judgements of Japanese speakers on the corresponding Japanese sentences, that should tell something about the linguistic relevance of the Knobe effect. For this purpose, we used two Japanese adverbs that should count as counterparts of English "intentionally". One is "itoteki ni", where "ito" (in Chinese character, "意図") means *intention* and "itoteki" means *intentional*. In fact, "ito" was a term created for translating "intention" or counterparts in other Western languages (cf. 金 2005). Thus "itoteki ni" is the standard and literal translation of "intentionally" into Japanese. However, because of this artificial origin, "itoteki ni" remains rather a formal expression, which is found mainly in formal texts, and not part of the colloquial expression in Japanese. In everyday conversation, the Japanese usually use "wazato" instead.[15]

A Japanese-English dictionary (GENIUS 2nd ed., TAISHUKAN) has four

[15] There is in fact another adverb, "koi ni (故意に)", which lies just in-between "itoteki ni" and "wazato", in the sense that it is not newly created recently like "itoteki ni", but is still a formal, rather than colloquial, expression. For comparison, however, we chose the two opposite ends of the continuum.

entries for “wazato”: 1) on purpose, 2) deliberately, 3) intentionally, and 4) purposely. The examples include: “He fell on purpose”, “I broke the vase on purpose”, “He deliberately neglected to call us”, “She gave me an intentionally vague answer”. As these examples suggest, this term is typically used when the agent does intentionally or on purpose what normally counts as *failure*. But this does not mean there is a semantic rule that prohibits the use of it in other contexts, or at least so it seems. One of the representative Japanese-Japanese dictionary (KOJIEN, 5th ed. IWANAMI SHOTEN), for example, gives no mention of the implication of the action being a failure, or any restriction in use, in the entry of “wazato”.[16] As the Japanese-English dictionary suggests, its primary sense is just *intentionally* (or its cognates).

Method:

We conducted the felicity judgment surveys on the use of all these adverbs, “intentionally”, “itoteki ni”, and “wazato”. 100 participants were recruited via Amazon M-Turk for the “intentionally” survey, and 100 participants were recruited via Lancer (a web service analogous to Amazon M-Turk) for each of the “itoteki ni” and “wazato” surveys. We eliminated two non-native English speakers from the data in the “intentionally” survey, ending up with 98 participants for “intentionally” (age M = 34.1, females 47%). There was no non-native Japanese speaker in the Japanese surveys, but due to the contingency of the system of Lancer, we were able to obtain more participants than originally planned, ending up with 106 participants for “itoteki ni” (age M = 38.6, females 52%), and 107 participants for “wazato” (age M = 36.1, females 45%).

[16] It gives four old Japanese uses, only one of which is the present sense of “intentionally”. Others are not in use or rarely found in contemporary Japanese.

Since we are interested in the concept of intentional action in the Knobe effect here, in order to collect the data of the linguistic judgements we used the following sentences.

HARM: X intentionally/itoteki ni/wazato harmed the environment.

IMPROVE: X intentionally/itoteki ni/wazato improved the environment.[17]

We used "improved" rather than "helped" because in Japanese there is no natural counterpart of "help" that has the environment as its object. *Kaizen suru* (or its post-tense form *kaizen shita*) is the standard translation of English "improve" (or "improved"), which we used for Japanese translations of IMPROVE.

We asked, in addition to HARM and IMPROVE, felicity judgments on the following sentences.

TUMBLE: X intentionally/itoteki ni/wazato tumbled.

BREAK: X intentionally/itoteki ni/wazato broke the vase.

STEAL: X intentionally/itoteki ni/wazato stole the purse.

SAVE: X intentionally/itoteki ni/wazato saved the life of Y.

KILL: X intentionally/itoteki ni/wazato killed Y.

IGNORE: X intentionally/itoteki ni/wazato ignored Y.[18]

TUMBLE, BREAK, and IGNORE were meant to be paradigmatic examples of the use of

[17] In the actual sentences presented to participants, "X" was replaced by "John" in the English survey, and "Taro" in the Japanese surveys.
[18] Similarly, "X" was replaced as described in the last footnote, and "Y" was replaced by "Betty" in the English surveys and "Hanako" in the Japanese surveys.

"wazato". Note that, TUMBLE is morally neutral and BREAK and IGNORE are only morally bad when done intentionally. So let us call these *Neutral*, or N for short. STEAL and KILL are meant to be of the same type as that of HARM, as morally bad behavior. Call these *Bad*, or just B. SAVE is meant to be of the same type as that of IMPOVE, as morally good behavior. Call these *Good*, or G. Thus, in sum, we conducted three surveys asking participants to judge the felicity of in total 8 sentences using "intentionally", "itoteki ni", and "wazato", respectively. Note also that the Japanese sentences are all perfectly grammatical for both *itoteki ni* and *wazato*, or at least no less grammatical than sentences with "intentionally" in English.

As for the order of sentences actually presented to participants, participants were divided into two groups, and in both groups the first two sentences were TUMBLE and BREAK, which are paradigmatic examples of "wazato", and then in one group they were presented 1) STEAL, 2) HARM, 3) SAVE, 4) KILL, 5) IGNORE, and 6) IMPROVE, while in the other group, after the first two sentences, the order was reversed, starting from 6 to 1.

After each sentence, participants were asked to answer how they found the use of the adverb in the sentence in question, by choosing one out of the following three options: 1) correct and natural, 2) not wrong but unnatural, and 3) wrong.[19]

Our initial hypothesis was that, in all three surveys IMPROVE, or sentences in G in general would be judged unnatural or wrong, while in the case of "wazato", HARM or sentences in B in general would be judged unnatural (but not wrong) since it looks too obvious that such bad actions are done *wazato*, and therefore adding "wazato" to them

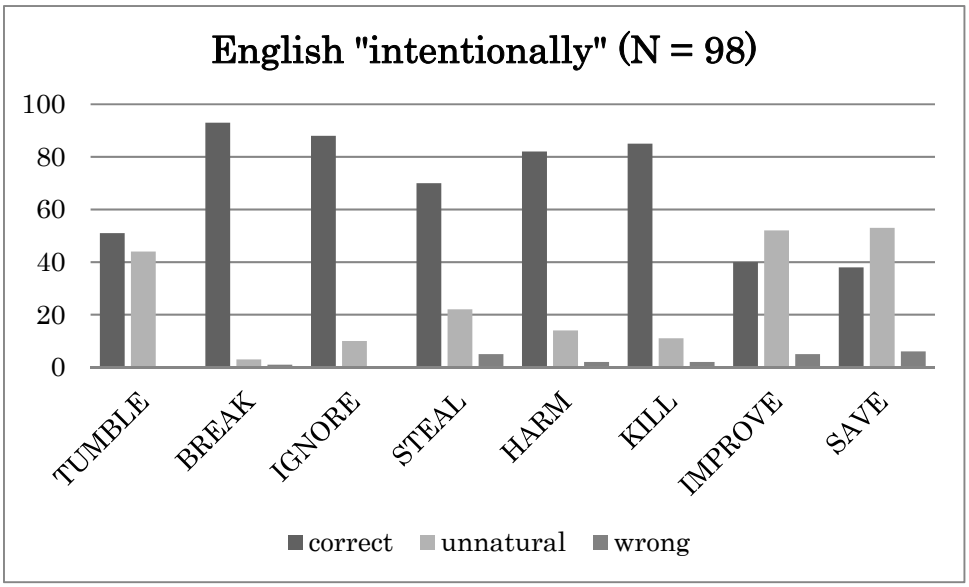
[19] In Japanese, 1) *Tadashiku shizen*, 2) *Machigai de wa nai ga, hushizen*, and 3) *Machigai*.

sounds just redundant.

Results:

1) English "intentionally"

We found more than half infelicity judgments (either unnatural or wrong) in sentences of G (61% for SAVE and 59% for IMPROVE), which share almost identical patterns, and nearly half (46%) infelicity judgments in TUMBLE, and the vast majority judged all others "correct and natural." Binomial analysis (sign test) showed that only the "correct and natural" answer of TUMBLE, SAVE, and IMPROVE failed to be significantly more than other answers. (Throughout three surveys we found no significant order effect.)



<Figure 2: English "intentionally">

This may be a *pragmatic* asymmetry in the sense that despite the large effect size, only ignorable minorities judged uses in SAVE and IMPORVE "wrong".

To determine whether grouping the sentences into Neutral, Bad, and Good better predicts the answer given by a given participant than without that grouping, we

conducted logistic regression analysis using the dichotomy "correct and natural" or otherwise (unnatural or wrong).[20] The likelihood-ratio test (log-likelihood statistic) showed that model N + B + G (where N, B, G are predictor variables and the "correct and natural" answer was the outcome variable) was a significantly better fit than the model TUMBLE + BREAK + IGNORE + … + SAVE (χ^2 (5) = 41.1, $p < 0.0001$), though the latter itself was a significantly better fit than the null model ($\chi^2(7) = 174.8$, $p < 0.0001$).[21] However, the former model was not significantly better fit than the model that does not distinguish N and B, that is, (N or B) + G (χ^2 (1) = 0.073, p = 0.79).

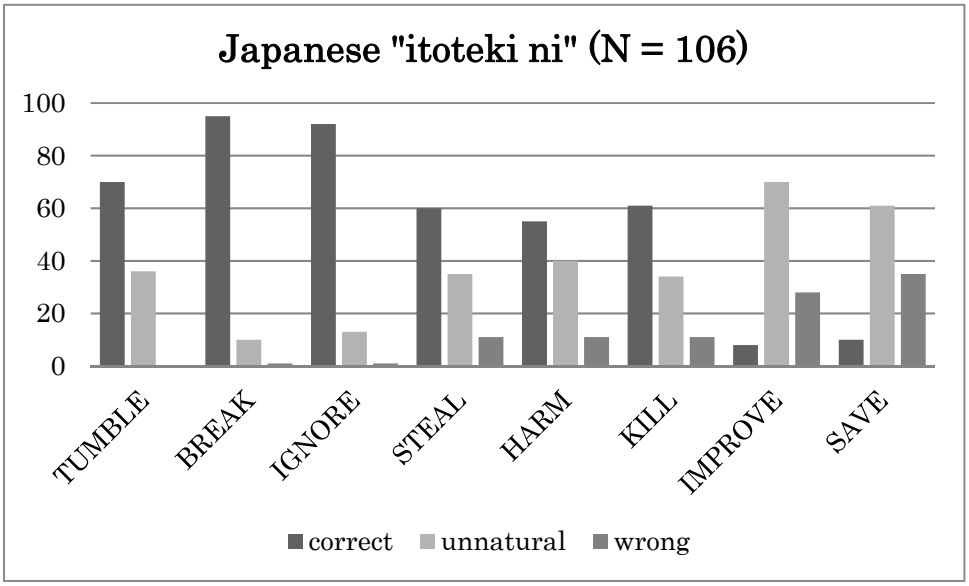
2) Japanese "itoteki ni"

Here we observed similar patterns in N-sentences (TUMBLE, BREAK, and IGNORE) on the one hand, and G-sentences (IMPROVE and SAVE) on the other, albeit significantly more infelicity judgments than those in the result of "intentionally". However, as binomial analysis showed that only the "correct and natural" answer of TUMBLE, BREAK, and IGNORE was significantly more than other answers, B-sentences (STEAL, HARM, and KILL), which also share a nearly identical pattern with each other, constituted an independent (third) group.

---

[20] Note that the decision to use this dichotomy is obviously *post hoc*. However, this is not *ad hoc* like the dichotomy "not wrong but unnatural" or otherwise (correct or wrong). We did not use the dichotomy correct (natural or unnatural) or "wrong" for it obviously has no difference between Neutral, Bad, and Good, which is why this is a post hoc analysis. See also the next footnote.

[21] One might think that, we should conduct ordinal regression analysis here. However, it is questionable that we can take "correct and natural", "unnatural", and "wrong" as *ordered* in this way, rather than categorical distinction. Besides, the present dichotomy makes perfect sense, as pointed out in the last footnote. Indeed, ideally, we could use the interaction model N*(TUMBLE+BREAK+IGNORE) + B*(STEAL+HARM+KILL) + G*(IMPROVE+SAVE). We avoided these methods just for the sake of simplicity, especially given the already extremely good fit of the present models (N + B + G).



<Figure 3: Japanese "intoteki ni">

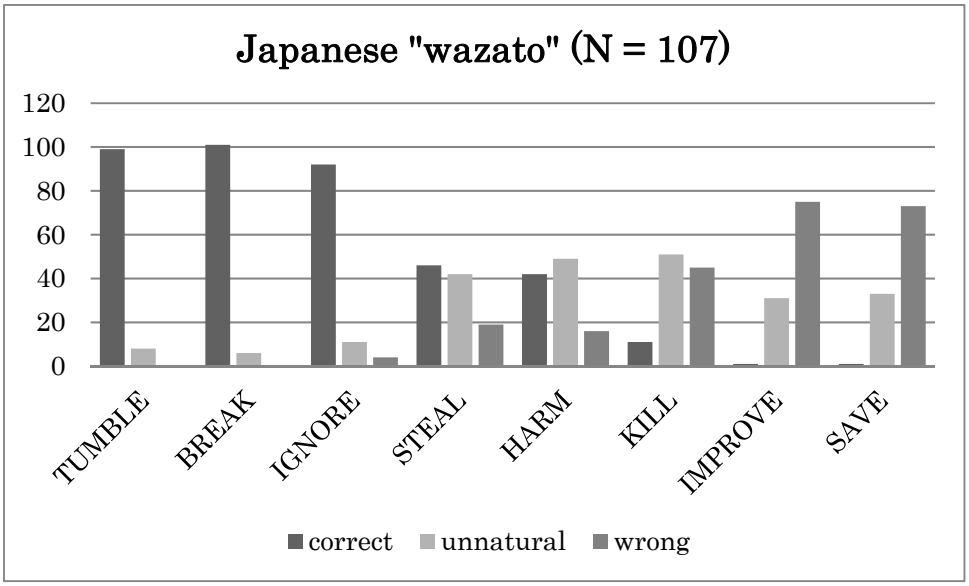
The same logistic regression analysis as in the "intentionally" survey (using the same dichotomy there) showed that the model N + B + G was significantly better fit with the data than both the base model TUMBLE + BREAK + IGNORE + … + SAVE (χ^2 (5) = 22.7, $p < 0.00039$) and the model (N or B) + G (χ^2 (1) = 48.5, $p < 0.0001$).[22]

3) Japanese "wazato"

Again almost the same pattern as that of "itoteki ni" is observed, where binomial analysis showed that only the "correct and natural" answer of TUMBLE, BREAK, and IGNORE was significantly more than other answers, though this time significantly more participants judged G-sentences "wrong" ($p < 0.0001$ for IMPROVE and $p = 0.0002$ for SAVE, two-tailed sign test). The result of KILL was however unexpected for its low rate of "correct and natural" answer (only 10% compared to 43% of STEAL and 39% of HARM),

[22] The latter two models themselves were significant (the base model: χ^2 (7) = 323.5, $p < 0.0001$, the model (N or B) + G: χ^2 (1) = 252.3, $p < 0.0001$).

and its high rate of "wrong" answer (42% compared to 18% of STEAL and 15% of HARM). The difference between KILL and STEAL, and the one between KILL and HARM, were both strongly significant (for the former, χ^2 (1) = 32.9, $p < 0.0001$, for the latter, χ^2 (1) = 32.0, $p < 0.0001$, two-tailed Chi-square Test, while effect sizes are both V = 0.39, according to Cramer's V, which is small if V = 0.1, medium if V = 0.3, and large if V = 0.5).



<Figure 4: Japanese "wazato">

Again, logistic regression analysis with the same dichotomy showed that the model N + B + G was significantly better fit with the data than both the base model TUMBLE + BREAK + IGNORE + … + SAVE (χ^2 (5) = 41.1, $p < 0.0001$), and the model (N or B) + G (χ^2 (1) = 267.8, $p < 0.0001$).[23]

[23] Again, the latter two models themselves were significantly better fit than the null hypothesis (the base model: χ^2 (7) = 608.0, $p < 0.0001$, the model (N or B) + G: χ^2 (1) = 299.0, $p < 0.0001$).

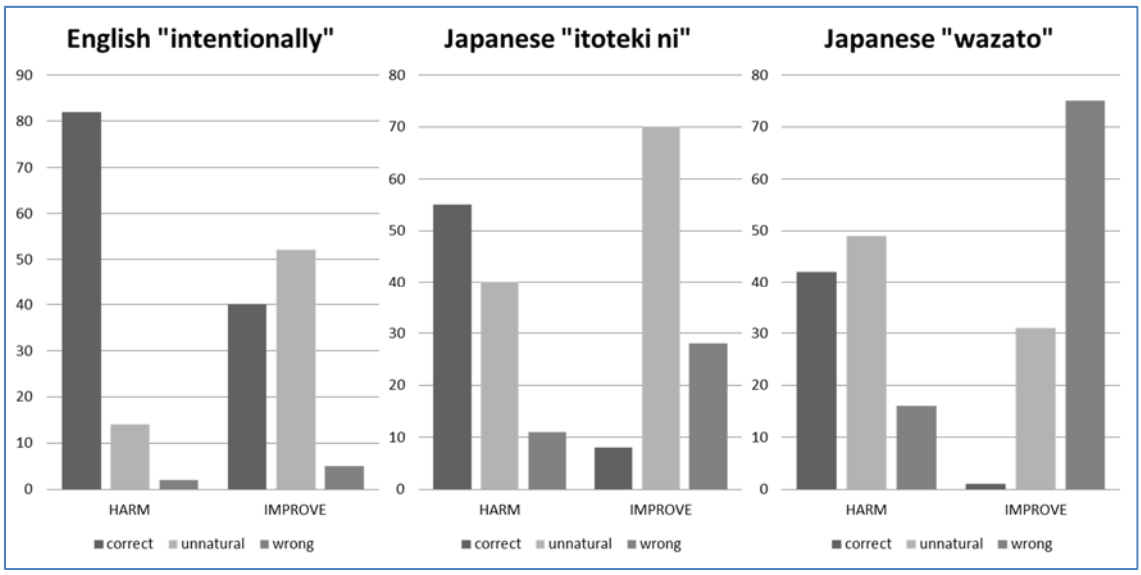
4) Comparison of three surveys

To compare three results, we should first note the striking similarity across the data of three adverbs. In particular, almost identical patterns found throughout three surveys between STEAL and HARM on the one hand, and SAVE and IMPROVE on the other, indicate that participants clearly took each pair to belong, respectively, to the same type.

There were two anomalies we found in the results: 1. TUMBLE in the "intentionally" and "itoteki ni" surveys, and 2. KILL in the "wazato" survey.[24] However, the results were otherwise almost exactly what had been expected. In particular, we found a *quasi*-Knobe effect between HARM and IMPROVE in all three surveys. Their differences are strongly significant, with large effect sizes (for all three, $p < 0.0001$, two-tailed Chi-square test, and df = 1, V = 0.44 for "intentionally", V = 0.49 for "itoteki ni", and V = 0.62 for "wazato", where the effect size is Cramer's V).

The comparison of these three linguistic Knobe effects shows a clear pattern:

[24] In the case of TUMBLE, the infelicity may come from the fact that it is simply not so easy to intentionally (or itoteki ni) tumble unless, for example, you are a trained comedian. In the case of KILL, there may be something special about the murder over and above its being morally bad. For example, there may be some special (semantic or pragmatic) connection between "wazato" and killing, so that "wazato" in KILL is especially infelicitous.



<Figure 5: Comparison of three "linguistic Knobe effects">

The "severity" of infelicity judgments of them is ordered as

"intentionally" < "itoteki ni" < "wazato",

for both HARM and IMPROVE (which is also the same as the order based on effect size).[25] The fact that "itoteki ni" is placed in-between "wazato" and "intentionally" in both HARM and IMPROVE is interesting. Since "itoteki ni" is not among the colloquial expressions in Japanese, and given its origin, it is natural to expect that its use follows (or even is parasitic on) the use of "intentionally". However, there is also a significant difference between "intentionally" and "itoteki ni" in the felicity judgments ($p < 0.0001$, two-tailed Chi-square test, for both HARM and IMPROVE, with effect sizes $V = 0.34$ for HARM, and $V = 0.44$ for IMPROVE).[26]

[25] The difference between "itoteki ni" and "wazato" in HARM failed to be significant ($p = 0.16$, two-tailed Chi-square test), but otherwise all other differences in both HARM and IMPROVE turned out to be strongly significant ($p < 0.0001$).

[26] One may therefore think here that the use of "itoteki ni" is also affected by "wazato",

3. Cross-linguistic difference and semantic/pragmatic account of the moral asymmetry

If we focus on the cross-linguistic difference of the results in the last section, it should be first noted, as for B-sentences, that even in the Japanese surveys only a minority judged, for example, HARMs to be *wrong* (10% in "itoteki ni" and 15% in "wazato"). One may therefore think that their respective (cross-linguistic) differences from HARM in "intentionally" are merely pragmatic, being effects of conversational maxims like *be informative*, or *don't be redundant*. Even so, however, since there no analogous pragmatic effect was observed in "intentionally", this difference, or why in English adding "intentionally" in these cases is *not* redundant, requires explanation. One hypothesis is that, there is a difference of view on what counts as guilty or blameworthy between Japanese and Americans. For example, stealing, harming the environment, and killing are already blameworthy for Japanese whether or not it is intentional, and therefore it does not have any point in adding either "itoteki ni" or "wazato", whereas for Americans intentionality plays a crucial role in the moral evaluation of such actions. If so, this is a cultural and conceptual, rather than psychological (in the sense of how the mind works) explanation.

On the other hand, the fact that G-sentences were generally judged *wrong* in "wazato" seems to show more clearly a cross-linguistic semantic difference from English counterparts in G. This does not mean, however, that Japanese people would judge "Taro *wazato* improved the environment" to be *false* (or even accept "Taro did NOT *wazato*

---

as well as "intentionally". To show this more rigorously might require surveys with within-subject design using bilinguals. But the point here is that, if this is a correct account, the effect in a survey with some term might be a result of the effect of some concept captured by some other term if a (possible) survey with the latter term has a larger effect size than the original survey (we shall come back to this issue in the final section).

improved the environment"), even when they were told that Taro improved the environment as his main purpose.[27] One might therefore think that the moral asymmetry even within the "wazato" survey is still merely pragmatic. But this assumes the semantic/pragmatic distinction based on *truth-condition*. We do not have to criticize truth-conditional semantics here,[28] but the distinction between "not wrong but unnatural" and "wrong" should be best captured as pragmatic/semantic distinction at least according to the *use theory of meaning*. Given such a theory, since there is nothing *syntactically* wrong with the use of "wazato" in sentences of either G or B, the most natural explanation of what is (not just *unnatural* but) *wrong* there should be semantic, or violation of a sematic rule (though this does not rule out the psychological explanation in terms of underlying cognitive processes either, as a supplementary account).[29]

Whether we call it semantic or not, however, the use of "wazato" suggests at least the existence of a concept according to which there is a moral asymmetry in its application, and we can easily imagine a language in which the moral asymmetry of this sort is wholly semantically encoded by some expression(s) as a matter of linguistic rule. In that case, the speakers of this language would find the Knobe effect rather trivial (which could be investigated by traditional conceptual analysis through the method of

---

[27] Rejecting such a possibility may be based on a principle that connects the correct application of a term with the truth of a sentence involving it, like "intentionally" is correctly applied to S's φ-ing if and only if "S intentionally φs" is true, which seems plausible enough.

[28] Alas, it turned out that there is a large difference in the use of truth predicates in different languages due to the difference of sensitivity to moral-political factor in the utterance (**** *manuscript*). But if so, the truth-conditional content itself should also vary from language to language.

[29] According to Bach (1997), the three major distinctions between semantic/pragmatic distinction are: 1. linguistic meaning vs. use, 2. truth-condition vs. non-truth-conditional meaning, 3. context independence vs. context dependence. Though all of them beg the question against some view of semantics, the present distinction is especially relevant to 3.

cases), with few of them regarding it as a *psychological* phenomenon. Such a possibility should be intelligible even for English speakers, as manifested in the apparently pragmatic asymmetry in the "intentionally" data, even though it is still possible that the effect found in English here reflects the semantic fact of "intention" or "intentionally" (for example, it has conventional implicature that the relevant action is not morally good[30]).

But if the effect is wholly pragmatic, Knobe was at least right about *English* in that the moral asymmetry in intention-attribution is not to be explained by conceptual analysis of *intentional action*, if it is not semantically encoded by the English term "intention" and its cognates.[31] However, he may not be right about other languages on this score. His and his colleagues' successful account in PLK (2015) in terms of the underlying cognitive processes, which does not refer to the linguistic factor (let alone the possibility of linguistic diversity), may have been applicable only to English speakers, being a rather contingent fact of English.

In any case, since the strongly significant *intra*-linguistic difference (for G-sentences) between "itoteki ni" and "wazato" (where cultural and psychological factors are controlled) must ultimately be traced back to some semantic difference between them, the concepts captured by these two expressions should be different, and therefore, even if the difference between "intentionally" and "wazato" was explained wholly pragmatically in terms of cultural-psychological difference, we would then still have the

[30] Note that, unlike *conversational* implicature, conventional implicature is semantic meaning (cf. Potts 2007, sec. 3).

[31] But if so, this suggests an unexpected return of the pragmatic account of the Knobe effect for English speakers (Adams & Steadman 2004a, 2004b, and Driver 2008a, 2008b), which is compatible with the psychological account but has long been regarded as refuted (cf. Knobe 2010, sec. 4.2). We need to take the pragmatic account seriously again at least in the case of English speakers.

cross-linguistic semantic difference between "intentionally" and "itoteki ni" anyway.[32]

One consequence of this (together with the argument from linguistic diversity) is that, we cannot naively assume *the* concept of *intention* or *intentional action* behind the relevant expression(s) that is supposed to be independent of any particular language anymore. When we discuss intention, we cannot be free from the question of which language we are using, or even which expression(s) among others within that language. Thus, the linguistic relevance of the Knobe effect is hard to deny.[33]

4. Linguistic default in the moral asymmetry of intention attribution

Given this linguistic relevance, we may take the results of section 2 as providing data of what we may call the *linguistic default*, which each sentence has relative to moral valence of the relevant action or event mentioned in the sentence, but is independent of any particular background or context, like how that action or event is caused and the details (in particular, the mental state) of the agent mentioned in it.[34, 35] Then, the linguistic diversity we have observed here suggests that the linguistic default may vary with language (or the concept captured by the relevant expression(s) in it). If so, the language-relative concept of intentional action captured by the relevant expressions in each language (or at least in English and Japanese) plays a role in producing the moral

[32] See also the same implication of two Japanese knowledge verbs, mentioned at footnote 13.
[33] I would like to thank Josh Knobe for suggesting me to clarify this point.
[34] Compare this with the notion of default in section 5.2 of Knobe (2010). It is ordinary people's reference point based on their conception of alternative possibility in each case, relative to which the agent's con/pro attitude is assessed. This is supposed to be determined by people's psychology, rather than linguistic norms (like our linguistic default).
[35] If so one might think that it could be investigated by traditional conceptual analysis through the method of cases. But even if that is true, mere conceptual analysis cannot determine the degree of the felicity of the kind we report here.

asymmetry, though this does not rule out the accompanying psychological effect.

We may then naturally wonder how this cross-linguistic difference affects the standard Knobe effect with the original vignettes of Knobe (2003a). For supplementary data, we have indeed conducted a survey using the Chairman Case in Knobe (2003a) with Japanese participants (see Appendix for the data). Unfortunately, there was no significant difference between "itoteki ni", "wazato", and "intentionally". This is however predictable, since given the linguistic defaults of three adverbs, we were supposed to have a difference of a larger effect size between HARM and IMPROVE for both of "itoteki ni" and "wazato", than for "intentionally", whereas the effect size of the latter in the original data of Knobe (2003a) was already quite large. Thus, this is a *ceiling effect* in the sense that, since the background story is expected to *enhance* or *strengthen* the effect, the cross-linguistic difference of the place of the linguistic default was overwhelmed by the effect of the vignette to become invisible. As Knobe himself favorably quotes from Cushman (2014), "it is often the case that big effects are best explained by a combination of many separate smaller effects" (PLK 2015, p. 41, sec. 7.2). What Knobe discovered in his 2003a was that this asymmetry gets particularly stark when the relevant event occurred as a side effect of the agent's causing the main effect. Since we found infelicity judgments even in the application of the relevant term to actions that are hard to think to have happened as mere side effects (like saving a life of another person), the asymmetry is *all the more* expected in its application to side effects (where the agent's attitudes toward them is not clear).[36]

[36] But if the context of use can strengthen the moral asymmetry already existing at the linguistic level, it may also *reduce* or *eliminate* the asymmetry. It might be possible to move the place of the linguistic default and render people judge SAVE and IMPROVE to be correct and natural even in the "wazato" survey, by providing an appropriate background story. If this happens, then *that* part of the effect is truly a psychological process over and above the linguistic/conceptual effect.

As suggested in the last section, some languages may encode the asymmetry semantically, while other languages may capture it only pragmatically. But still others may have nothing to do with morality, showing utterly no pragmatic, let alone semantic, asymmetry whatsoever. The linguistic default of this language is therefore totally independent of moral valence. If we find such a language, however, it is interesting to ask whether the speakers of it show the Knobe effect in response to the original vignettes of Knobe (2003a). If we can still find the Knobe effect there, it is truly and purely a psychological process for them. If we find, surprisingly though, no such effect with them at all, we may be able to give even a purely *linguistic* (semantic or pragmatic) account of the Knobe effects found in speakers of all other languages. In any case, we need to examine the place of the linguistic default for each language in order to say whether and to what extent the Knobe effect is a psychological phenomenon for the speakers of that language.

If we do not like such actual and possible linguistic diversity and the consequent linguistic relevance of moral asymmetry we have discussed so far, we should rather take the concept of intentional action to be something *normative*, independent not only of any particular language, but also of any cognitive processes. But this brings us back to the point of Knobe's criticism of the traditional approach of conceptual analysis. Meanwhile, as long as we are willing to learn from empirical facts, the account of the moral asymmetry in question cannot be exhausted by how the mind works (as opposed to how we use the relevant term). For, even if the difference found in the data of the felicity judgments between "intentionally" and "wazato" here could be wholly explained by the difference of the underlying cognitive processes behind the judgments, the very cognitive processes should also be influenced by the relevant concepts, as we pointed out in section

1. (Indeed, we may think that the difference in the data between “itoteki ni” and “wazato” is the instance of such case.)

Even though there are various psychological accounts of the moral asymmetry of intention-attribution, therefore, they are still compatible with the significance of the present linguistic approach, and it has even a virtue of avoiding possible unnecessary complexities due to the details of the vignettes in the standard questionnaire research, revealing the linguistic defaults of relevant sentences.

5. Conclusion and Future Work

The aim of this paper has been to defend and promote the present simple linguistic approach to the Knobe effect, and we have shown that the Knobe effect, or the moral asymmetry of intention-attribution in general, can be reproduced through the mere felicity judgement on the relevant sentences, with large effect sizes.

Then we can easily see that there are two (mutually compatible) dimensions of extending the scope of the present linguistic investigations on the Knobe effect.

1) surveys with terms other than “intentionally” and its cognates.
2) surveys with languages other than English and Japanese

As for 1, we already know that there are analogous moral asymmetries for some other expressions like “know”, “caused”, “allowed”, etc. if we use the relevant vignettes (as mentioned in section 1). However, it is unlikely that we will find the similar asymmetric patterns for these expressions *without* any vignette. At least, in the case of English, it seems that the asymmetry of “intentionally” has the largest effect size. Then,

in general, the term with the largest effect size may be understood as capturing and expressing the *central concept* of moral asymmetry, in the sense that the asymmetries found in the standard questionnaire research with vignette using other expressions are rather *derivative* of it, in the sense that, in Japanese, "wazato" is the central concept and the moral asymmetry of "itoteki ni" may be a derivative of it.[37] Such central concept may even vary from language to language, and we can observe whether or to what extent this pattern is shared universally through the surveys of type 2. For example, in some language the causation judgment may have the largest effect size of moral asymmetry while there is only small effect size for intention attribution.

In any case, these two dimensions will provide us with a matrix that will systematically reveal the facts about the general phenomena of the moral asymmetry in different languages. Such a study, together with the standard surveys using the vignette, will contribute to the study of the dynamism between concepts and the cognitive processes, which will in turn support the supplementary picture of experimental philosophy in a broad sense, in that conceptual analysis together with the empirical study of concept and the mainstream experimental philosophy research focusing on the underlying cognitive process supplement each other. Given this picture, or to examine it more closely, the linguistic approach to the Knobe effect of the present kind is worth further pursuing anyway.[38]

References:

Adams, F., & Steadman, A. (2004a). Intentional action in ordinary language: Core

[37] See footnote 26 above.

[38] This work was supported by JSPS KAKENHI (C) Grant Number JP26370010.

concept or pragmatic understanding? *Analysis*, 64, 173–181.

Adams, F., & Steadman, A. (2004b). Intentional actions and moral considerations: Still pragmatic. *Analysis*, 64, 268–276.

Alexander, Joshua (2012), *Experimental Philosophy: An Introduction*. Polity Press.

Buckwalter, W. & Phelan, M. (2014). Phenomenal Consciousness Disembodied. In *Advances in Experimental Philosophy of Mind*, edited by Justin Sytsma, New York: Bloomsbury, 45–73.

Cushman, F. (2014). What scientific idea is ready for retirement? Big effects have big explanations. Retrieved from <http://edge.org/response-detail/25508>.

Davidson, D. (1973) "Radical Interpretation", in *Inquiries into Truth and Interpretation,* Oxford: Clarendon Press (1985), pp. 125-140.

Davidson, D. (1975) "Thought and Talk", in *Inquiries into Truth and Interpretation,* Oxford: Clarendon Press (1985), pp. 125-140.

Driver, J. (2008a). Attributions of causation and moral responsibility. In W. Sinnott-Amstrong (Ed.). *Moral psychology: The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 423–439). Cambridge, MA: MIT Press.

Driver, J. (2008b). Kinds of norms and legal causation: Reply to Knobe and Fraser and Deigh. In W. Sinnott-Amstrong (Ed.). *Moral psychology: The cognitive science of morality: Intuition and diversity* (Vol. 2, pp. 459–461). Cambridge, MA: MIT Press.

金光林(2005), 近現代の中国語、韓国・朝鮮語における日本語の影響―日本の漢字語の移入を中心に―新潟産業大学人文学部紀要　第１７号, 111-238.

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

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation.

Philosophical Psychology, *16*(2), 309-324.

Knobe, Joshua (2007) "Experimental Philosophy and Philosophical Significance," *Philosophical Explorations* 10, 119-122.

Knobe, J., & Nichols, S. (2008). An experimental philosophy manifesto. *Experimental philosophy*, edited by J Knobe and Shaun Nichols, 3-14. New York: Oxford.

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

Knobe, J. (2016). Experimental philosophy is cognitive science. In Sytsma, J., & Buckwalter, W. (Eds.). (2016), 78-96.

Machery, E. (2016). Experimental Philosophy of Science. In Sytsma, J., & Buckwalter, W. (Eds.). (2016), 475-490.

Mizumoto, M. (*forthcoming*) "Know" and Japanese Counterparts: "Shitte-iru" and "Wakatte-iru", M. Mizumoto, S. Stich, E. McCready, and J. Stanley (ed.) *Epistemology for the Rest of the World*, Oxford University Press.

Mortensen, K. & Nagel, J. (2016). Armchair-friendly experimental philosophy, In Sytsma, J., & Buckwalter, W. (Eds.). (2016), 53-70.

Murray, D. & Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research* 88: 434– 467.

Nahmias, Eddy, Stephen G Morris, Thomas Nadelhoffer, and Jason Turner. (2006). Is Incompatibilism Intuitive? *Philosophy and Phenomenological Research* 73: 28– 53.

Nadelhoffer, T. (2005) Skill, luck, control, and folk ascriptions of intentional action. *Philosophical Psychology* 18:343–54.

Nadelhoffer, T., & Nahmias, E. (2007). The past and future of experimental philosophy. *Philosophical Explorations*, *10*(2), 123-149.

Nichols, S., & Knobe, J. (2007). Moral responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Nous* 41: 663– 685.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30-42.

Potts, C. (2007). Into the Conventional - Implicature Dimension. *Philosophy compass*, *2*(4), 665-679.

Reuter, K., Phillips, D. and Sytsma, J. (2014). Hallucinating Pain. In *Advances in Experimental Philosophy of Mind*, edited by Justin Sytsma, New York: Bloomsbury, 75–99.

Robinson, B., Stey, P., & Alfano, M. (2015). Reversing the side-effect effect: the power of salient norms. *Philosophical Studies*, *172*(1), 177-206.

Stich, S., & Tobia, K. (2016). Experimental philosophy and the philosophical tradition. In Sytsma, J., & Buckwalter, W. (Eds.). (2016), 5-21.

Strickland, B., Fisher, M., Knobe, J., & Keil, F. (2015). Syntax and intentionality: An automatic link between between language and theory-of-mind. *Cognition*, *133*(1), 249-261.

Sytsma, J., & Buckwalter, W. (Eds.). (2016). *A companion to experimental philosophy*. John Wiley & Sons.

Sytsma, J., & Machery, E. (2013). Experimental philosophy. *Encyclopedia of philosophy and the social sciences*, 319-321.

Appendix: Japanese data of the survey with the original Chairman Case

We omit here the demographic details of the participants, but they were recruited through Lancer with exactly the same procedures as in the main text. The results are summarized in the following table with the data of "intentionally" taken from Knobe (2003a) for comparison.

	HARM	IMPROVE
itoteki ni	86.0% (N=50)	11.8% (N=51)
Wazato	76.8% (N=56)	12.5% (N=56)
intentionally	82% (N=39)	23% (N=39)

As for HARM, there was no significant difference (Fisher's exact test) between the results of "itoteki ni" and "wazato". The statistical power ($1-\beta$) of detecting the difference between these two in HARM for effect size of 0.3 and this sample size (N=106) was 0.87.

In the case of IMPROVE, even if we doubled the sample size of "intentionally" to 78 (while keeping the percentage constant, which we assumed 9/39), neither the difference with "itoteki ni" nor "wazato" was significant (Fisher's exact test).

Overall, we found no significant difference between three adverbs (even after doubling the sample size of "intentionally") either in HARM or IMPROVE (Chi-square test), which was predictable as discussed in section 4.