Pinning down intentions

Bridget Copley

CNRS/Paris 8

Language clearly cares about intentions. Yet both in our theories of syntactic structure and our theories of meaning, the role of intentions remains somewhat inchoate. In this talk, I present the state of the art on intentions at the syntax-semantics interface, discuss reasons why intentions have been left out in the cold, and show how representing intentions in causal models can allow us to gain empirical coverage with respect to linguistic data involving intentions.

NAÏVE RATIONALIZATION

John Schwenkler*

I

One way of giving *rationalizing explanations* of intentional action—that is, explanations of someone's doing something in terms of their reasons for acting—is with words like 'want', 'think', and 'intend'. For example:

(1) My reason for flipping the switch was that I wanted to turn on the light.
(2) I am easing the jib because I think that will stop the main from backing.
(3) I am pulling weeds because I want a beautiful lawn.
(4) I turned left at the fork because I wanted to get to Katmandu.
(5) James went to church with the intention of pleasing his mother.

Since the 1960s, the orthodox position in philosophy of action has had two parts: first, that statements like these are the *primary* way of rationalizing action; and second, that words like 'want' and 'intend', as they figure in these explanations, should be understood as referring to the agent's *mental states*.

In the second part of *Life and Action* (=*LA*), Michael Thompson makes an argument that this position should be inverted. According to Thompson, the primary form of the rationalization of action does not appeal to mental states, but rather is *the explanation of one action by another*, where the act that is explained "might be said to be a part, phase, or 'moment'" of the one that explains it (*LA*, 86). For example (see *LA*, 85):

(6) I'm pulling this cord because I'm starting the engine.
(7) I'm cutting these wires because I'm repairing a short circuit.
(8) I'm crossing Fifth Avenue because I'm walking to school.
(9) I'm breaking these eggs because I'm making an omelet.

At least on the surface, there is no outward movement on display in these sentences. Instead, the form of explanation they involve is that of a part by a larger whole.

Following Thompson, let's call the pattern exhibited in our first group of sentences, **sophisticated rationalization**, and the pattern exhibited in our second group, **naïve rationalization**. Here are the skeletal forms that we'll treat as the paradigms of each:

(SR) … A-ing because … intend(s) (to) B
(NR) … A-ing because … B-ing

Further, and more radically, Thompson advances the "hypothesis" (*LA*, 131) that the proper understanding of words like 'want' and 'intend', as they appear in rationalizing explanations like (1)-(5), is not as referring to inner states that are distinct from outer activity, but rather as

* Department of Philosophy, University of Illinois Urbana-Champaign. Email: jlschwenkler@gmail.com.

describing the same kind of "imperfective presence" (*LA*, 131) that's described in statements like (6)-(9), in which imperfective aspect is used to describe the progress of an ongoing but uncompleted activity.

A crucial step in Thompson's defense of this hypothesis is his attempt, at the end of chapter 8 of *LA*, to describe "a form of life and thought" in which sophisticated rationalization "is simply unknown" (*LA*, 93):

> Among such agents, all of the work of straightforward rationalization is effected by means of the rationalization connective combined only with the categories of ordinary event consciousness. The more "sophisticated" forms of straightforward rationalization can then be depicted as arising from this rustic state of things in a series of stages … (*LA*, 92)

My focus in this paper is on working through this thought-experiment. Specifically, I'll follow Thompson in trying to understand how the process of "sophistication" through which the naïve agents add to the expressive resources of purely (NR)-type rationalizing statements could involve no more than applying "the categories of ordinary event consciousness" to the imperfective process-descriptions that naïve rationalizations appeal to.

II

Here are six components of our ordinary thought about events that will prove relevant to understanding the structure of naïve rationalization.

First, events are understood to "unfold" in a way that *takes* time. A mark of this is our use of imperfective aspect ('is … A-ing') to describe events as they take place.

Second, and relatedly, many kinds of event can be *uncompleted*. This possibility will be there whenever we can say of an entity that it *was A-ing but did not A*.

Third, events can have *proper parts or phases*: e.g., if a stone rolls from α to γ, and β is a point lying between them, then *in* rolling from α to γ the stone rolls from α to β and from β to γ.

Fourth, events can *depend* on one another, as expressed in our use of connectives like 'because'.

Fifth, sometimes the description of what is happening at a time refers to a *wider context* than what can be seen to be going on at that very moment. Sometimes this is called the "broadness" of the progressive aspect.

Finally, events can be "imminent", or describable as things that are *going to* happen, or that *are happening* at a future point in time. Note that imminence is not the same as bare futurity: as with incompletion, a thing may *have been going to* do something that it *did not do*.

III

The question we are considering is whether the categories of ordinary event-consciousness, as

outlined in the last section, could be sufficient for the rationalization of human action. Before we can answer this question, we need to reflect on the way that these categories are transformed in their application to the practical domain.

The answer I'll give is inspired by Anscombe's argument in the opening sections of *Intention*. It can be put very directly by saying that *a rationalizing statement must be able to be given first-person expression*, as exemplified in the original statements (6)-(9). We can see by noting several points of contrast between those statements and the following:

(10) Sue is driving erratically because she is looking at her phone while she drives.
(11) You are sweating because you are exercising so vigorously.
(12) I am gaining weight because I am drinking a milkshake every day.

Each of (10)-(12) has the surface form of (NR), but none of the three offers a rationalizing explanation. Aside from the possibility of overt appeal to intention, how does this difference find expression in what we say?

First, each of (10)-(12) describes something that could be happening without the agent's having any idea of this. To see this, contrast (10) with a third-person counterpart of (6):

(13) She's pulling the cord because she's starting the engine.

Unlike (10), what (13) says can't be true, or at least can't be a true rationalizing explanation, if the agent doesn't know that she's pulling the cord, or that this is why she is doing it.

Second, each of them describes something that the agent could sensibly be *informed* that she is doing, and could come to know this as a consequence of being so informed. Here we may contrast (11) with (14):

(14) You are pulling the cord because you are starting the engine.

Unless it's in the mouth of a psychoanalyst, (14) could only be a restatement of something that the speaker has been told by the person she is speaking to. But (11) could be the advice of a personal trainer, on the basis of which the agent comes to understand both that she's sweating, and why.

Third, while (12) gives us an example of a non-rationalizing statement that's in the first person, it's read naturally as the expression of something the agent *discovered*, perhaps through the advice of a dietitian. But none of (6)-(9) can be read in this way.

IV

Suppose, then, that we have a group of speakers who can rationalize action only through a suitably transformed version of the simple event-consciousness described in section II. Would this be enough for them to have, and give expression to, a robust concept of intentional action as something that can be "rationalized", and so *done for reasons*?

I hope we have seen how these naïve agents could articulate and grasp rationalizations of the sort shown in (6)-(9). Further, we can see how the understanding of events as potentially "imminent" could ground the expression and attribution of future intention, where a person's present action is understood through something else that she is *going to do*. E.g.:

(14) I am packing a bag because I am taking a trip tomorrow.
(15) I am buying wood because I am going to build a birdhouse next week.

How, though, are they supposed to think about cases where a person simply *isn't* doing, or going to do, the thing in terms of which their action would be rationalized? E.g., consider this naïve re-statement of our original (1):

(16) I am flipping the switch because I am turning on the light.

If the light is broken, then what (16) says can't be true—at least, not unless a person's description of her own intentional action is allowed to be true by fiat. So while the agent might assert (16) sincerely, *we* could not endorse (17), but at most (18) or (19):

(17) She is flipping the switch because she is turning on the light.
(18) She is flipping the switch because she intends (wants) to turn on the light.
(19) She is flipping the switch because she thinks it will turn on the light.

Likewise for a statement like (14), in a case where we know that the speaker isn't going to take the trip after all, say because the travel agency has canceled it. There, we might say something like (21), but not (20):

(20) She is packing a bag because she is taking a trip tomorrow.
(21) She is packing a bag because she plans to take a trip tomorrow.

Notably, the same phenomenon recurs in the first-person form, in any case where the agent is uncertain about what she's doing or going to do. Thus, e.g., one who's unsure if the light is working will assert not (16) but (1), and one who's unsure if the trip will happen will assert not (14) but (22):

(22) I am packing a bag because I hope to take a trip tomorrow.

These cases present a residual puzzle: how can words like 'intend', 'want', 'think', 'plan', and 'hope' be read as describing the "imperfective presence" of an action, given that the action in question is not present at all?

In the last part of my talk, the solution I'll suggest is that we can understand how these sophisticated rationalizations relate to naïve ones by analogy with the way that a person can be seen as acting in light of belief rather than knowledge.

How to Perform a Nonbasic Action

Mikayla Kelley

Forthcoming in *Noûs*

Some actions we perform "just like that" without performing any other actions.[1] Think: raising your arm, wiggling your finger, or winking. These are called *basic actions*. Other actions—the *nonbasic actions*—are performed by performing other actions. The actions by which one performs a nonbasic action are the *means* by which one performs the nonbasic action. Think: voting by raising your arm, illuminating a room by flipping a switch, or walking across the street by taking multiple steps in sequence.[2]

The topic here is the kinds of means we take to nonbasic action. Most are committed to something like the following picture:

> (**Necessity of Constitutive Means**) Where φ is a nonbasic action, if an agent φs by intentionally ψ-ing, then ψ—the agent's means to φ-ing—constitutes φ.

That is, the nearly ubiquitous view is that we only take *constitutive* means to nonbasic action. Kieran Setiya spells out this assumption explicitly:

> "It is a necessary truth about nonbasic action that if one does A by doing B, doing B is a constitutive not productive means to doing A: It is an instance of doing A or a part of the process of doing A, not just a prior cause that makes it happen. That is why, although I can cause myself to blush by dropping my trousers in public, I do not count as blushing intentionally, not even as a nonbasic action, when I do so".

I'll argue that the ubiquitous acceptance of Necessity of Constitutive Means is a mistake, one which is symptomatic of an incorrect understanding of basic features of intentional action and has led to an underestimation of our agential capacities. In particular, I argue that we must make room for *productive* means to nonbasic action, where a productive means causes rather than constitutes that which it is a means to. Think: sneezing by looking up at the sun, kicking one's prosthetic leg by pressing a button, or laughing by bringing to mind a funny joke. I will argue that in all three cases one performs a nonbasic action—sneezing, kicking, or laughing—and one does so using productive means.

One upshot of the general reflections on nonbasic action here is progress made with regards to understanding the extent of our mental agency and, in particular, our capacity to form beliefs intentionally (what is called *doxastic voluntarism*). Indeed, we often take productive means to mental movements and once we properly understand that doing so is a way to perform a nonbasic action, we see that we have more mental agency than some have thought; we intentionally judge by seeking out evidence, intentionally decide by weighing pros and cons, and intentionally recall a fact by mentally cycling through associations, even though in all three cases one takes a merely productive means to the relevant mental movement.

Additionally, in order to make room for productive means, we have to reflect on basic features of intentional action including control, purposefulness, and agent participation. I will show that it is only through an overly narrow understanding of each that we rule out productive means to nonbasic action. I defend instead the existence of what might be called a *pluralism of manifestations* of control, purposefulness, and agent participation: there are multiple ways to instantiate or realize each feature in action. Thus, a second upshot of our investigation into the means we can take to nonbasic action is an illumination of the nature of numerous building blocks of intentional action more broadly.

[1] I use 'action' and 'intentional action' interchangeably.

[2] This is a standard distinction between basic and nonbasic action that focuses on *teleological* basicness and nonbasicness.

The plan for the talk is as follows. I present the principle of *Causal Closure*, which is a set of sufficient conditions for intentional action in terms of using a productive means. The principle is as follows:

(**Principle of Causal Closure**) Assume that

(1) an event M is a movement of an agent's under the description $\ulcorner\varphi\urcorner$;

(2) the agent intends to φ and this intention persists until the completion of M;[3]

(3) an event N is an intentional action of the agent's under the description $\ulcorner\psi\urcorner$;

(4) the agent uses ψ as a productive means to φ;[4]

(5) and N non-deviantly causes M.[5]

Then the event M is an intentional action of the agent's under the description $\ulcorner\varphi\urcorner$.

After presenting Causal Closure, I describe a motivating example which will be the basis of my argument for Causal Closure and against Necessity of Constitutive Means: Ananya kicking her prosthetic leg by pressing a button. I'll argue that there is no relevant distinction between Ananya and Jonny who kicks his leg in the usual way. The crux of my argument will involve confronting what seems to be the biggest obstacle to Causal Closure: its being inconsistent with a conception of action that requires a sufficient level of control to be possessed at the time of action. After making room for productive means to nonbasic action, I'll conclude by showing how doing so allows us to respond to skepticism about our capacity for robust mental agency.

[3] Assuming that the intention persists until the completion of M rules out cases where the agent dies, goes into a coma, etc., prior to M (though it is worth noting that the first condition might be enough to rule out these cases).

[4] I'll explain what this requires in the talk.

[5] As Donald Davidson has taught us, we need to be careful when appealing to causation in conditions for action, for not any causal relation will suffice for action. There are *deviant* causal relations that do not entail agency. So, following Davidson, the notion of non-deviance is used here as a placeholder for a precisification of the kind of causal relation between the productive means and the movement to which the productive means is aimed at that suffices for the movement to be an intentional action.

Situations, intentions, and locality

Kenyon Branan and Rob Truswell

AIL4, January 2024

Truswell (2011) argued that intentions indirectly constrain A′-movement, the grammatical dependency found in *wh*-questions, relative clauses, and many other constructions. Two pieces of evidence were: (a) A′-movement out of rationale clauses, which describe purposes, is often possible, even though rationale clauses are adjuncts and adjuncts generally prohibit movement (1a); (b) A′-movement out of bare present participial adjuncts is grammatical only if one of the two verb phrases in question describes a nonagentive eventuality (compare (1b–c)).

(1) a. What did you come here [to talk about __]?
b. What did you sit around [whistling __]?
c. *What did you work [whistling __]?

In this earlier work, I proposed a semantic constraint on A′-movement, the Single Event Condition, and argued that intentions featured in the delimitation of single events.

In this talk, we revisit this link between intentions and movement. We believe that my earlier account was approximately correct in spirit, but incorrect in most of the details. We first demonstrate that the semantic units to which movement is sensitive are situations, not necessarily corresponding to single events, and motivate a constraint on movement dependencies, the **Imbrication Requirement**, that relies on the individuation of situations. Secondly, we develop a syntactic implementation of the Imbrication Requirement that draws on treatments of noncanonical switch reference systems, in which certain grammatical morphemes reflect information about the individuation of situations. These morphemes are key to an interpretable, compositional implementation of the account we develop here, and allow us to draw indirect connections between intentions, which are implicated in the semantics of switch reference, and locality constraints on movement.

1 What is imbrication?

Consider a sentence containing two heads X and Y, each introducing a situation variable, respectively s_x and s_y. In principle, these two variables could be interpreted independently of each other, in a logical form along the lines of (2). Of course, in a coherent discourse, there would be *some* indirect relation between s_x and s_y, established via times, worlds, or rhetorical relations, but no relation is directly established between the situation variables — there is no clause of the form $R(s_x, s_y)$ in (2).

(2) $\exists s_x \exists s_y \ldots (P(\ldots s_x \ldots) \wedge Q(\ldots s_y \ldots) \wedge \ldots)$

Alternatively, s_x could be interpreted as *part of* s_y, yielding a logical form like (3).[1]

(3) $\exists s_x \exists s_y \ldots (P(\ldots s_x \ldots) \wedge Q(\ldots s_y \ldots) \wedge s_x \sqsubseteq s_y \wedge \ldots)$

In (3), a relation is established directly between s_x and s_y, using the part-of relation $\sqsubseteq$. We call this configuration **imbricational**, and that in (2) **independent**.

[1] Our analysis is also largely compatible with an implementation where s_x and s_y are both part or a third, larger situation, s_{xy}.

We will show that many complex sentences are ambiguous between an imbricational and an independent LF, but movement often disambiguates in favour of the imbricational LF. We call this the **Imbrication Requirement**: movement requires imbrication. A main goal of this talk is to state this requirement more clearly, and to motivate it.

We do not believe that it is possible to give necessary and sufficient conditions to diagnose imbricational LFs, although contingent relations such as causation, enablement, and intention clearly favour them. In this talk, we will focus on the disambiguating effect of movement, which already allows us to illustrate the effects of the Imbrication Requirement.

Consider the interpretation of temporal adjuncts with *when* (parallel arguments can be made with *after* and *before*).[2] *When* has two interpretations, an independent one which relates times and an imbricational one which relates situations. The imbricational interpretation is apparent in (4), from Moens & Steedman (1988): on this interpretation, there are apparently no purely temporal requirements imposed by *when* (the architect's plans precede the building, possibly by years; the using materials is part of the building; the solving the traffic problems follows the completion, possibly playing out over years). What is constant across these three examples is the imbricational relation holding between *P* and *Q*: (a–c) are interpreted as *part of* the building of the bridge.

(4) [When they built the 39th Street bridge], ...
   a. a local architect drew up the plans.
   b. they used the best materials.
   c. they solved most of their traffic problems.

The independent interpretation of *when* can be forced by adding a measure phrase describing how precisely the situations described by *P* and *Q* overlap temporally. Examples are *roughly*, *approximately*, and *exactly*. Adding any of these measure phrases to (4) removes the imbricational reading, leaving an interpretation where *P* and *Q* describe independent, temporally proximate situations.

(5) [(Approximately/Exactly) when they built the 39th Street bridge], ...
   a. a local architect drew up the plans.
   b. they used the best materials.
   c. they solved most of their traffic problems.

The Imbrication Requirement tells us that movement out of the *when*-clause forces the former, imbricational readings and disallows the latter, independent readings. This means that movement is incompatible with measure phrases like *approximately* or *exactly*, an otherwise quite unexpected interaction.

(6) Snakes like this, you need to be careful [(*precisely) when you touch __].

2 Imbrication and A-movement

Similar effects can be found across many types of movement, whether A- or A′-, out of adjuncts or complements, in a range of languages. The *when*-clauses in Section 1 illustrate the case of A′-movement out of adjuncts, although there are many more examples. Here, we demonstrate the effects of the Imbrication Requirement in two cases of A-movement.

A-movement out of adjuncts: Hebrew possessor raising Example (7), from Landau (1999), has two interpretations. Either ('dull lecture') Gil was present in the lecture, and the lecture sent him to sleep; or ('lazy student') Gil slept through his alarm and missed the lecture. Different continuations ('... it was so boring' / '... and that's why he wasn't there') disambiguate in favour of one or the other reading.

[2] See work in progress by Caroline Heycock, Elise Newman, and Rob Truswell.

(7) Gil yašan be-zman ha-harc'a
G. slept in-time the-lecture
'Gil slept during the lecture.' (Landau 1999, p. 19)

In our terms, the 'lazy student' reading is independent, while the 'dull lecture' reading is imbricational. When this sentence is combined with possessor raising out of the adjunct (which Landau shows to be an instance of A-movement), only the imbricational reading survives, so a continuation which requires the independent reading is infelicitous.

(8) Gil yašan le-Rina be-zman ha-harc'a, #ve-laxen hu lo higia
G. slept to-Rina in-time the-lecture, and-therefore he not come
'Gil slept during Rina's lecture, #and that's why he didn't come.'

A-movement out of complements: Japanese scrambling Japanese scrambling within a finite clause shows properties typical of A-movement, such as feeding scope or binding relations. Most instances of scrambling across a finite clause boundary do not feed scope or binding, and are therefore analysed as A′-movement (Saito 1992).

However, scrambling out of obligatory control clauses can have A-properties, for instance feeding binding in (9b).

(9) a. *[Soko-no sotugyoosei-ga] [PRO mittu-izyoo-no daigaku-ni syutugansi-yoo to] sita
it-GEN graduate-NOM three-or.more-GEN university-DAT apply-will COMP did
b. [Mittu-izyoo-no daigaku-ni] [soko-no sotugyoosei-ga] [PRO __ syutugansi-yoo to] sita
three-or.more-GEN university-DAT it-GEN graduate-NOM apply-will COMP did
'Their$_i$ graduates tried to apply to three or more$_i$ universities.' (Takano 2010)

This is unlikely to follow from structural properties of OC complements, because they look like any other finite complement clause. In fact, the same effect is also in evidence when the embedded subject is bound *pro*, not PRO (Funakoshi 2015).

(10) a. *[Soko$_i$-no raibaru-gaisya-no syain-ga] [*pro*$_i$ mittu-izyoo-no kaisya-ni$_i$
it-GEN rival-company-GEN employee-NOM three-or.more company-DAT
oobosurusuru-tumorida to] itta
apply-will COMP said
b. [mittu-izyoo-no kaisya-ni$_i$] [soko$_i$-no raibaru-gaisya-no syain-ga] [*pro*$_i$ __
three-or.more company-DAT it-GEN rival-company-GEN employee-NOM
oobosurusuru-tumorida to] itta
apply-will COMP said
'Employees of their$_i$ rival companies said that *pro*$_i$ will apply to three or more companies$_i$.'

Unbound *pro* does not license scrambling in the same way. Japanese cross-clausal A-scrambling therefore has an interpretive requirement which is formally similar to the Imbrication Requirement, in that movement eliminates certain interpretive options while preserving others. However, it plays out not in terms of contingent relations such as causation or enablement, but in terms of referential dependencies between individual arguments (see also Grano & Lasnik 2018).

3 Making sense of imbrication

We think that the Imbrication Requirement is informative about the interpretation of chains formed by DP-movement. We note, firstly, that all the examples we have gathered in Sections 1–2 of the Imbrication Requirement, involve movement of DPs. This is straightforwardly true of the A-movement cases in Section 2. However, all the A′-movement cases that that we are aware of, like that in Section 1, are the kind of restricted A′-dependencies known as 'A′-binding' in Cinque (1990) and as 'B-dependencies' in Postal (1998).

We assume that DP-traces are interpreted along the lines suggested by Fox (2002) and Elbourne (2005): a mechanism that Fox calls 'trace conversion' converts a copy of the form (11a) into a definite description of the form (11b), where s′ is a variable over situations.

(11) a. (Det) Pred
b. [the s′] [Pred y.(y=x)]

We can then reformulate the Imbrication Requirement as a condition on the interpretation of DP-traces, as follows (see also Gluckman 2018).

(12) **Imbrication Requirement** (2nd version)
The head and foot of a chain formed by movement of a DP must be interpreted with respect to situations that stand in an imbricational requirement.

Intuitively, the Imbrication Requirement is a requirement that the head and foot of the chain pick out the same object — an object that uniquely satisfies the same description, in a closely related situation. Without the Imbrication Requirement, this isn't guaranteed, because nothing forces the situation variables in different links in the chain to be bound in this way.

4 Imbrication and switch reference

We have assumed that many structures are ambiguous between implicational and independent interpretations. As a straightforward compositional interpretation of this ambiguity, we posit two operators, which are inserted at the left edge of the relevant clauses and other constituents. One of these operators (which we call $\mathsf{Link}_{\mathrm{SS}}$, for reasons which will become apparent) binds a situation variable in its complement and asserts that this situation stands in an imbricational relation to a c-commanding situation variable. The other operator ($\mathsf{Link}_{\mathrm{DS}}$) binds a situation variable and asserts that it does *not* stand in an imbricational relation. The effect of the Imbrication Requirement is then to enforce the use of $\mathsf{Link}_{\mathrm{SS}}$ along the relevant movement path, and prohibit the use of $\mathsf{Link}_{\mathrm{DS}}$, for interpretability reasons: $\mathsf{Link}_{\mathrm{DS}}$ makes the movement chain uninterpretable.

A virtue of this compositional implementation is that it reuses semantic devices independently motivated by McKenzie (2012) in his analysis of **noncanonical switch reference**. Canonical switch reference markers are morphemes which appear near the edge of a subordinate clause, and indicate whether that clause's subject is identical, or nonidentical, to the subject of a superordinate clause (respectively known as same-subject, or **SS**, and different-subject, or **DS**, markers). In *non*canonical switch reference, the SS marker appears even though the two subjects in question are disjoint. An example, from Kiowa, is in (13).

(13) *Context: a letter-writing campaign*

Kathryn	gʲæ-gú?	**gɔ**	Ester=al	gʲæ-gú?
K.	3S.3PL-write	**and.SS**	E.=also	3S.3PL-write

'Kathryn wrote a letter, and Ester wrote one too.' (McKenzie 2012, p. 159)

In McKenzie's analysis, *gɔ* in (13) indicates that Ester's letter-writing forms part of a larger situation, together with Kathryn's letter-writing. In other words, *gɔ* is an overt morphological indicator of an imbricational relation, of the sort that we have argued is required for interpretation of DP-traces.

5 Intentions and locality

The situation-based analysis described in this abstract has several advantages over the event-structural approach of Truswell (2011):

- The implementation of the Imbrication Requirement sketched above is embedded in a maturing understanding of the semantics of A′-movement, building on the Fox/Elbourne approach to the interpretation of traces, McKenzie's analysis of switch reference, and Gluckman's work on traces and situations. The Imbrication Requirement is also *interpretable* in the terms of that framework, in that we can propose explanations of why such a constraint might hold. In contrast, the Single Event Condition from Truswell (2011) was a stipulation, coming from nowhere.
- The empirical scope has broadened significantly, in that we can address various facts involving different types of movement, out of adjuncts and complements, finite and nonfinite, in a way that was beyond the scope of my previous work.

But the most relevant advantage, for the purposes of this workshop, is that the link to noncanonical switch reference provides an independent source of evidence into the relationship between intentions and situations, of the sort implicated in the Imbrication Requirement. Returning to the letter-writing example (13), it is not sufficient for the licensing of the SS morpheme *gɔ* that Kathryn and Ester are engaged in similar activities — McKenzie shows that a DS morpheme would surface instead if that were the only link. Rather, the important point is that Kathryn and Ester share a common purpose: the campaign is a series of letter-writing events unified by their intention.

This provides direct grammatical evidence that intentions and goals delimit the semantic units with reference to which movement is constrained. It also reinforces our arguments that the semantic units in question are not events. However, it is also slightly unsettling, in that the way in which intentions feature in the discussion of noncanonical switch reference is different in many respects from the way in which it features in previous discussions of A′-movement. The next step in this line of research is then a process of optimistic triangulation: hoping that the ideas sketched here are on the right track, and using them to look for parallels between the way in which intentions license noncanonical SS morphemes, and the way in which they license noncanonical movement operations.

References

Cinque, Guglielmo. 1990. *Types of $\overline{A}$-dependencies*. Cambridge, MA: MIT Press.

Elbourne, Paul. 2005. *Situations and individuals*. Cambridge, MA: MIT Press.

Fox, Danny. 2002. Antecedent-contained deletion and the copy theory of movement. *Linguistic Inquiry* 33(1). 63–96.

Funakoshi, Sayaka. 2015. *A theory of generalized pied-piping*. University of Maryland, College Park dissertation.

Gluckman, John. 2018. *Perspectives on syntactic dependencies*. University of California, Los Angeles dissertation.

Grano, Thomas & Howard Lasnik. 2018. How to neutralize a finite clause boundary: Phase theory and the grammar of bound pronouns. *Linguistic Inquiry* 49(3). 465–499.

Landau, Idan. 1999. Possessor raising and the structure of VP. *Lingua* 107(1–2). 1–37.

McKenzie, Andrew. 2012. *Austinian situations and switch-reference*. University of Massachusetts, Amherst dissertation.

Moens, Marc & Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics* 14. 15–28.

Postal, Paul. 1998. *Three investigations of extraction*. Cambridge, MA: MIT Press.

Saito, Mamoru. 1992. Long distance scrambling in Japanese. *Journal of East Asian Linguistics* 1(1). 69–118.

Takano, Yuji. 2010. Scrambling and control. *Linguistic Inquiry* 41(1). 83–110.

Truswell, Robert. 2011. *Events, phrases, and questions*. Oxford: Oxford University Press.

Agentivity, animacy, prototypicality and specialized meaning

Malka Rappaport Hovav
The Hebrew University of Jerusalem

Beth Levin
Stanford University

Many English verbs are *variably agentive*: they are found with either agentive or non-agentive subjects, as illustrated with *push* in (1).

1. a. Pat pushed the stroller.
   b. The current pushed the boat.

For most English verbs the prototypical use – the first use that comes to mind – is an agentive use such as in (1a), most likely because of the salience of the animate entities that are agents. A reasonable hypothesis would be that there is not much difference in meaning and grammatical behavior between agentive and non-agentive uses of variably agentive verbs. Both sentences with *push*, for instance, involve the exertion of a force away from the entity denoted by the subject despite the difference in their subjects' agentivity. Yet, although this prediction holds for of *push* and many other verbs, it does not hold of all verbs, including a set of variably agentive verbs which includes *sweep, bake,* and *teach* that are the focus of this talk. These verbs have an agentive use which is not only taken to be the prototypical use of the verb, but it also is obligatorily agentive, showing a narrowing of the meaning found with other uses of the same verb, which are not obligatorily agentive. In particular, such necessarily agentive uses involve the lexicalization of a routine activity of an agent that represents a specialized instance of the event encoded by the verb's basic meaning.

The distinctive properties of these verbs can be brought out by considering the first sentences that come to mind with the verbs *sweep, bake*, and *teach*: they might look like those in (2), and speakers of English would take them to instantiate prototypical uses of these three verbs.

2. a. Pat swept the floor.
   b. Tracy baked cookies this morning
   c. The substitute taught the class today.

In these sentences, the verb is necessarily agentive. For instance, the subject in (2a) cannot be replaced by a natural phenomenon, as in (3a); further, if a comparable sentence is modified with *accidentally*, as in (3b), the interpretation is that the location swept isn't the intended one, not that the action of sweeping itself is accidental.

3. a. *The wind swept the floor. (cf. 1b)
   b. Pat accidentally swept under the table.

In this respect, *sweep* patterns like an obligatorily agentive verb such as *assassinate,* as shown in (4), which has to be interpreted as a case of mistaken assassination.

4. The sniper accidentally assassinated the king's bodyguard.

But *sweep* has a broad range of uses that do not show necessary agentivity: it can be found with both inanimate subjects and animate subjects, which may or may not act intentionally, as in (5).

5. a … when the branch of the tree swept the window.
   b. The waves swept the deck.
   c. The storm swept the debris out of the valley.
   d. Pat (accidentally) swept the harp strings with her fingers.
   e. Kelly (accidentally) swept the papers off the desk.
   f. Gina (accidentally) swept her hands against the freshly painted fence.
   g. Ash swept through the streets.

Concomitantly, such uses of *sweep* don't suggest themselves as prototypical instances of the verb. The prototypical, necessarily agentive use shows other semantic restrictions besides obligatory agentivity. It must involve manipulating a broom over a floor-like surface, as shown in (6), contrasting with the other uses, which lack these restrictions: (5d) involves the use of fingers and (5e) has a desk as the surface.

6. a. *Pat swept the kitchen floor with a shovel.
   b. Pat swept the deck/patio/walk/yard.
   c. *Pat swept the desk/the window/the wall/the book.

We claim that goal-oriented human activities have a tendency to get lexicalized, deriving specialized, narrowed senses of otherwise variably agentive verbs, whose basic sense is unspecified for agentivity. Although the specialized sense retains the same semantic core as the basic sense, because of its association with a goal-oriented activity of humans, this sense is taken to be the verb's 'prototypical' sense in that it represents the prototypical activity named by the verb. Thus, from the point of view of 'building verb meaning' the prototypical sense of the verbs in question – *sweep, bake, clean, wash,* and *teach* – is **not** the basic sense.

It is a special property of *sweep* and its kin that what is taken to be their prototypical sense reflects a specialized meaning, which is necessarily agentive. In contrast, the prototypical activity named by many *systematically variably agentive verbs* may involve an agent, but there is no reason to take their prototypical instances to represent a specialized sense, as with the verb *push* in (1). As another example, consider the verb *topple* in (7). The agentive (7a), with a human subject, represents a prototypical instance of toppling, but it is not describing a different type of situation from (7b), with a natural phenomenon subject. Further, agentivity with this verb is always defeasible: (7c), for instance, could felicitously be continued with *they did it by mistake!*

7. a. The kids toppled the Lego tower with glee.
   b. The hurricane toppled the TV tower.
   c. Our activists were cleared of criminal damage for toppling a statue of slave trader Edward Colston … (*Mirror* (Nexis) 6 January)

Agentive uses of systematically variably agentive verbs then do not bring with them a specialized meaning. The agentivity of animate subjects of these verbs is attributed to a pragmatic inference applying to animates (Van Valin & Wilkins 1996). For *sweep, push,* and *topple* the prototypical uses are agentive (cf. the Idealized Cognitive Model of an event of Croft 1991, DeLancey 1984, Lakoff 1987, Langacker 1987). However, for *sweep* and comparable verbs such as *bake* and *teach* the prototypical and non-prototypical uses differ in crucial lexical properties. This is not so for *push, topple,* and many other verbs.

The *sweep* case study elaborated

We illustrate our proposal that inherently agentive uses of variably agentive verbs involve a specialized meaning with an extended analysis of the English verb *sweep* before turning to some other verbs. We argue that there is a basic sense of *sweep* that underlies all its uses and is unspecified for agentivity; it simply involves an entity moving over a surface while maintaining contact with it. This sense brings together a wide range of situation types with subjects of varied ontological types, as illustrated in (5). The event structure in (8) represents the grammatically relevant elements of meaning of the verb which determine its argument realization options.

8. basic-*sweep*: "x moves across a surface y and x imparts a force to y via contact".

We show that the argument realization options associated with the basic meaning of *sweep* come from allowing either the movement predicate or the imparting of force predicate in (8) to determine argument realization. When the motion predicate determines argument realization, two related structures are derived by established principles of argument realization: (i) an unaccusative+PP structure, the syntactic structure which expresses motion along a path, as in (5g), and (ii) a causativized version of (i), which yields a transitive+PP structure, as in (5c-f). When the imparting force predicate determines argument realization, established principles of argument realization yield a transitive structure, as in (5a,b).

The prototypical use of *sweep* can be analyzed as involving a specialized sense that retains the semantic core of the basic sense (8) but is derived from it by saturating the variable x, requiring it to be a broom, as in (9).

9. broom-*sweep*: "x_{broom} moves across a surface y and x imparts a force to y via contact".

In this specialized sense, *sweep* then gets interpreted like those denominal verbs taking their names from instruments, such as *funnel, mop,* and *staple*. Such verbs must denote an activity representing the canonical use of the instrument (Kiparsky 1997). We show that this simple adjustment to the event structure (8) has wide-ranging consequences for argument realization and can explain the different argument realization options for the two senses of *sweep* (Levin & Rappaport Hovav 2022). In particular, we explain why the unaccusative+PP and transitive+PP frames are unavailable for the broom-*sweep* sense. Furthermore, because the canonical use of a broom represents a routine goal-oriented activity, only the broom-*sweep* sense allows unspecified object uses (e.g., *Sam swept this morning* must be interpreted as involving a broom), as expected since such uses are licensed when a verb describes a routine goal-oriented activity (Glass 2022). In this respect, *sweep* contrasts with *topple*: although, as mentioned, prototypical instances of *topple* also have an animate agentive subject, they do not describe a routine activity of an agent and, as illustrated in (10), *topple* does not allow unspecified objects.

10. *The toddler **topples** every time he builds a tower.

Instances of all instrument-based denominal verbs are interpreted as canonically performed activities involving the source instrument, as in (11). For instance, (11b) must be understood as involving a use of a funnel that fulfills its design purpose: the sand must be poured into the funnel. It cannot describe an event of pushing sand off a table and into a cup by moving a funnel in a 'sweeping' motion across the table.

11. a. I **mopped** the floor
    b. I **funneled** the sand into the cup.

The verbs *mop* and *funnel* provide evidence that the unspecified object frame is only available if an agentive activity is routinized: that is, it is always done in a specific way (Brisson 1994; Glass 2022; Mittwoch 2005). Mopping is such an activity, and the related verb has an unspecified object use; funneling is not such an activity, and the related verb lacks an unspecified object use.

12. a. I **mopped** all morning.
    b. ?I **funneled** all morning.

Moving beyond *sweep*: Other routine goal-oriented activities of agents are lexicalized

Sweep's specialized meaning derives from the lexicalization of 'broom'. But specialized meanings can arise independent of the lexicalization of an instrument. Many activities of agents tend to be performed in specific ways to fulfill particular goals, so that they have a tendency to become routinized; subsequently, special, narrowed interpretations of the relevant verb become licensed. To illustrate this, we present two further case

studies of non-denominal verbs that show that *sweep* represents a larger phenomenon: other verbs that can describe routinely performed activities may lexicalize a specialized sense. We discuss *bake* (Atkins, Kegl & Levin 1988) and *teach*, although we could make comparable arguments using the verbs *clean* (Levin & Rappaport Hovav 2014) and *wash* (in the grooming sense).

The meaning components common across instances of *bake* are a change of state that comes about through the application of heat. This meaning is found in unaccusative uses, as in (13a,b), and in transitive uses with both agentive and non-agentive subjects, as in (13c-e).

13. a. The potatoes **are baking** in the oven.
    b. The bricks **are baking** in the sun
    c. The sun **is baking** the creek bed.
    d. The potter **is baking** a dozen vases in the kiln.
    e. The chef **baked** some apples for brunch.

But *bake* has a narrower use to describe the agentive activity of making baked goods as in *Tracy baked cookies this morning*. This is what English speakers would consider the prototypical use of the verb. It is this narrower meaning that is associated with unspecified object uses of the verb. *Tracy baked this morning* can only be used if what is being baked is baked goods such as bread, cakes, or cookies; it cannot be used to describe baking vegetables or chicken; nor can it describe baking vases or other ceramics in a kiln even with a potter as the subject of the verb.

Turning next to the verb *teach*, teaching can take many forms: a person can teach a child to ride a bicycle or swim, a dog to beg, a new employee how to do their job, or an apprentice how to fix light fixtures. Furthermore, the subject of *teach* need not be agentive; the verb takes a range of subjects, as in (14).

14. a. This video **taught** me how to fix the light fixture.
    b. The sudden storm **taught** me to always close the windows before I go out.

But the prototypical event described by *teach* is classroom teaching, which we take to be a lexicalized sense reflecting a routine goal-oriented activity. In this sense, the verb has the hallmarks of such verb senses. The verb is obligatorily agentive in this sense, as shown by the interpretation of *Kim accidentally taught the class how to solve the first homework problem*, where what is accidental is what is taught and not the activity of teaching itself. It is also found with unspecified objects; for instance, *Kim taught this afternoon* must refer to classroom teaching and not, say, to Kim teaching her dog a new trick.

Animacy is the key to the lexicalization of specialized meaning: The *drown* case study

Abstracting away from the discussion of these three verbs, we propose that there is a regular process of lexical specialization of verb meaning that involves routine activities of agents. This specialization gives rise to the unspecified object frame with the relevant verbs. Such uses are generally taken to be prototypical instances of the action denoted by the verb. We propose that the prototypicality of certain agentive uses of verbs and the tendency for such uses to get lexicalized follows because they involve routine activities of animates. Evidence that animacy rather than agentivity is the key to such lexical specialization comes from the verb *drown*, which takes a patient argument. As we now show drawing on Rappaport Hovav (2017), with this verb the prototypical use, which manifests lexical specialization, is associated with a non-agentive but animate argument, i.e. the verb's patient.

The first use of *drown* that suggests itself – that is, its prototypical use – is that in (15), which involves an animate entity, the verb's patient, who dies due to immersion in water.

15. The boy *drowned* (?but the paramedics were able to save him before he died).

The parenthetical continuation in (15) shows that death is entailed in this use. However, Rappaport Hovav (2017) shows that generally this verb does not lexically encode the death of the patient, even when the patient is animate, as shown by (16).

16. … your mommy can … soap you [a dog] and **drown** you and dry you …
(http://dogvotional.blogspot.co.il/2010/04/; accessed 1/7/2024)

Nor does drowning have to involve water, as in (17a,b), or involve an animate entity, as in (17a).

17. a. The cake is **drowning** in icing.
    b. They **drowned** Natalia Portman in fabric to hide her pregnancy.

Rappaport Hovav takes these examples to reflect the basic meaning of *drown* and proposes that the event structure for this basic meaning is as in (18), which like *sweep*'s event structure involves two components.

18. basic-*drown*: "x **bears a spatial configuration** with respect to y such that y **covers** x"

As with basic-*sweep*, argument realization principles apply to either one of the bolded components of meaning, giving rise to either transitive or unaccusative/causative instances of *drown*, as in (16) and (17), respectively. We propose that the instances that have an entailment of death due to immersion in water as in (15) represent a lexicalized meaning that fixes the value of y in (18) to water, restricts x to animate entities, and entails x's death, as in (19).

19. specialized-*drown*: "$x_{animate}$ **bears a spatial configuration with respect to** y_{water} such that y **covers** x bringing about x's death"

This specialized meaning, which involves an animate entity, is again taken to be the prototypical meaning; however, unlike with *sweep, bake,* and *teach,* in this instance the specialized meaning involves a patient. Hence, this example shows that animacy is the key to what is taken to be the prototypical use of a verb.

Conclusion

In summary, we have shown that verbs whose prototypical use involves an agent typically do not lexically require an agent. However, variably agentive verbs sometimes develop a specialized agentive sense derived from a basic sense which is unspecified for agentivity via the lexicalization of a goal-oriented activity of an animate entity. If this activity is routinized, the verb may be found in the unspecified object construction in this sense. Given the nature of the activity and the salience of animate entities, this specialized sense then represents the prototypical use of the verb. However, the development of specialized senses involving prototypical uses is more fundamentally associated with animate entities, agents being one instance, although the more common one.

References

Atkins, B.T. et al. (1988) Anatomy of a verb entry, *International Journal of Lexicography* 1:84-126.
Croft, W. (1991) *Syntactic Categories and Grammatical Relations*, University of Chicago Press.
DeLancey, S. (1984) Notes on agentivity and causation, *Studies in Language* 8:181-213.

Glass, L. (2022) English verbs can omit their objects when they describe routines, *English Language and Linguistics* 26:49-73.
Langacker, R.W. (1987) *Foundations of Cognitive Grammar 1*, Stanford University Press.
Levin, B. & M. Rappaport Hovav (2014) Manner and result: A view from *clean,* in *Language Description Informed by Theory*, John Benjamins, 337-357.
Levin, B. & M. Rappaport Hovav (2022) Conventionalized agentive activities and compositionality, *QMUL Occasional Papers in Linguistics* 47.
Rappaport Hovav, M. (2017) Grammatically relevant ontological categories underlie manner/result complementarity, *IATL 2016*, MITWPL 86, 77-98.
Van Valin, R.D. & D.P. Wilkins (1996) The case for 'effector': Case roles, agents, and agency revisited, in *Grammatical Constructions*, Clarendon Press, 289-322.

The mental representation of causation explains Kraemer's puzzle

Tadeg Quillien, School of Informatics, University of Edinburgh

Suppose Joan wants to kill Bill. There is a lever in front of her: pulling the lever will randomly open one of ten boxes in the room where Bill stands. While box eight contains poison, all other boxes are empty. Joan pulls the lever: Luckily for her (but not for Bill), box eight opens, releasing the poison and killing Bill.

People have the intuition that:

1. Joan intentionally killed Bill,

But they tend to deny that:

2. Joan intentionally opened box eight.

This pattern in prima facie puzzling: Bill dies if and only if box eight opens, so pulling the lever raised the probability of both events equally. This case is an instance of Kraemer's puzzle: when an agent brings about X as a means to an end Y, philosophers and laypeople sometimes judge that the agent did Y intentionally but did not do X intentionally (Butler, 1978; Kraemer, 1978; Blumberg & Hawthorne, unpublished manuscript; Pavese and Henne, 2023).[1]

I argue that this pattern has a natural explanation: the Kraemer effect is what we should expect given what cognitive science has revealed about the folk concept of intentional action. A cue to what is happening is given by our intuitions about *causation*. Consider the following statements:

1'. Bill died because Joan wanted to kill him,

2'. Box eight opened because Joan wanted to open box eight.

Intuitively 1') seems much better than 2'). A natural hypothesis is that this asymmetry in causal attribution explains the corresponding asymmetry in intentionality attribution. That is, people are more likely to say that Joan intentionally killed Bill (compared to opening box eight) because they are more likely to think that wanting to kill Bill caused her to kill Bill.

There is in fact already extensive evidence that mental representations of causation are a core building block of the concept of intentional action. Therefore, our intuitions about causation provide a natural explanation to Kraemer's puzzle. In the following, I first briefly describe causalist accounts of intentional action and explain how they account for the Kraemer effect. I then discuss how these ideas relate to other recent explanations of the effect.

I) Intentional action and the mental representation of causation.

[1] The effect has also been demonstrated in morally neutral cases (Nadelhoffer, 2004; Butler, 1978; Pavese & Henne, 2023), but the asymmetry is most striking in those cases where the outcome is morally bad.

It seems strange to say that someone did X intentionally if their desires had no causal influence whatsoever on the fact that X happened. Accordingly, causalist accounts hold that causation is an important component of the concept of intentional action (e.g. Davidson, 1980). In cognitive psychology, a causalist account has recently been defended by Quillien & German (2021). Roughly speaking, they argue that an agent did X intentionally if the agent's attitude towards X (i.e. how much they want X to happen) caused X.

As an example, consider the famous **chairman** case:

Chairman. The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also [harm/help] the environment.' The chairman of the board answered, 'I don't care at all about the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was [harmed/helped]. (adapted from Knobe, 2003a)

Participants reading the 'harm' version of the story tend to agree with the statement:

3. The fact that the chairman does not care about the environment caused the environment to be harmed,

But participants in the 'help' condition tend to disagree with the statement:

4. The fact that the chairman does not care about the environment caused the environment to be helped. (Quillien & German, 2021).

According to the causalist theory, this asymmetry explains the classic asymmetry in intentionality attributions: people agree that the chairman intentionally harmed the environment (in the harm version), but disagree that he intentionally helped the environment (in the help version). The theory also correctly predicts many other empirical features of people's judgments, for example the fact that actions are seen as more intentional if the agent had control over the outcome.

As such, the causalist explanation of Kraemer's puzzle, outlined above, is independently supported by many empirical findings about the concept of intentional action. One might still complain that the causalist explanation simply replaces a mystery with another: we proposed that the asymmetry in intentionality attributions was due to an asymmetry in causal attributions, but where does this asymmetry in causal attributions come from?

Fortunately, the pattern of causal intuitions can be naturally explained in terms of psychological theories of how people represent causation. According to these theories, causal judgment involves counterfactual reasoning (Gerstenberg et al., 2021; Icard et al., 2017; Quillien & Lucas, 2023). Roughly, people judge that C caused E to the extent that, if C had not happened, then E would not have happened. So to a first approximation, when people judge:

1'. Joan wanting to kill Bill caused Bill to die,

people are evaluating the counterfactual:

1''. If Joan had not wanted to kill Bill, Bill would not have died.

Intuitively this counterfactual is true: if Joan hadn't wanted to kill Bill, she would not have pulled the lever, and Bill would have carried on living.

Similarly, when people judge:

2'. Joan wanting to open box eight caused box eight to open,

they are evaluating the counterfactual:

2''. If Joan had not wanted to open box eight, box eight would not have opened.

The truth of this counterfactual is less clear. For example, if Joan had wanted to open box three instead, she would still have pressed the lever, and this could have opened box eight.[2]

In sum, counterfactual theories of causation naturally predict that there will be an asymmetry in causal intuitions in a Kraemer-like case. And causalist theories of intentional action predict that this causal asymmetry will give rise to the Kraemer effect.

Next we discuss how these ideas relate to other recent explanations of the Kraemer effect.

II) Know-how and intentional action.

Pavese and Henne (2023) recently put forward an elegant explanation of the Kraemer effect. They point out that people tend to agree with:

1'''. Joan knows how to kill Bill,

But disagree with:

2'''. Joan knows how to open box eight.

This is of course the mirror of the pattern for intentionality judgments. Since there is independent evidence that judgments of know-how influence judgments of intentionality (Pavese, Henne & Beddor, 2023), Pavese and Henne argue that the asymmetry in know-how judgments naturally explains Kraemer's effect. They also provide empirical evidence, across many studies, that i) know-how and intentionality judgments strongly co-vary, ii) experimental manipulations of know-how have a strong effect on intentionality judgments.

I suggest that the know-how account is consistent with the causalist account. The key idea is that, when an agent doesn't know how to do X, it seems strange to say that their desire for X caused X. Consider the following prototypical case of an agent who does X without knowing how to:

[2] Evidence from cognitive psychology suggests that this counterfactual outcome (where the outcome of the lever pull is similar to the actual-world outcome) is particularly salient, given that counterfactual reasoning is biased toward situations that are similar to what actually happened (Lucas & Kemp, 2015; Quillien, Szollosi, Bramley & Lucas, 2023).

Bull's-eye. Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bull's-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. (Knobe, 2003b)

It seems strange to say: 'Jake hit the bull's eye because he wanted to'. Instead, we say that he hit the bull's eye because he got very lucky. This effect of know-how on causal attributions is a relatively straightforward prediction of contemporary cognitive models of causal judgment. These models feature a **robustness** condition: roughly, 'C caused E' requires that C would still have led to E even if background conditions had been slightly different (Lombrozo, 2010; for different ways of implementing the robustness condition see Icard et al., 2017; and Quillien, 2020; Quillien & Lucas, 2023).

This robustness condition fails to hold in **Bull's eye**: we can easily imagine situations where Jake wants to hit the bull's eye but fails. As such, we don't judge that his desire to hit the bull's eye is the main cause of his success.

In sum, the effect of know-how on intentionality judgments is consistent with a causalist account of intentional action: when an agent does not know how to do X, their desire for X fails the robustness condition and thus does not count as a strong cause of X. The question still arises of whether the effect of know-how on intentionality ascriptions could be distinct from the effect of causation. We can shed some light on this issue by looking at a case where know-how and causation are dissociated.

Consider a variant of **Bull's eye** where Jake is actually an expert sharpshooter, but happens to sneeze right as he makes the shot. The shot goes wild … By astonishing luck, the bullet still lands directly on the bull's-eye. Here, I have the intuition that:

5. Jake knows how to hit the bull's eye,

But I tend to disagree with:

5'. Jake wanting to hit the bull's eye caused him to hit the bull's eye,

5''. Jake intentionally hit the bull's eye.

In other words, intentionality seems to track causation instead of know-how in this case.

III) Probability-raising, alternative-sensitivity and intentional action

People are more likely to judge that an agent did X intentionally if X's action significantly increased the probability of X (Ericson et al., 2023). At first sight, a probability-raising account seems unable to explain the Kraemer effect, because pulling the lever raised the probability of both events equally.

However, Blumberg and Hawthorne recently argued that people might compute probabilities over different sets of alternatives when they reason about the means and the ends (Blumberg & Hawthorne, ms.). For example:

-when people think about opening box eight, they think about whether pulling the lever increases the probability that box eight will open, *relative to any other box*.

-when people think about Bill dying, they think about whether pulling the lever increases the probability of killing Bill, relative to not killing Bill.

Probability-raising accounts are consistent with a causalist theory. There is a well-known link between causation and probability-raising, as causes tend to increase the probability of their effects. So, one natural hypothesis is that manipulations of probability-raising affect intentionality attributions because they influence causal attributions.

Our theory must also account for Blumberg & Hawthorne's observation that intentionality ascription is sensitive to the nature of the alternatives that are raised in a context. Blumberg & Hawthorne demonstrate alternative-sensitivity by highlighting an effect of linguistic focus. Consider a variant of our main scenario where there are ten potential victims: if box one opens, Peter (and only Peter) dies, if box two opens Mary dies,…, if box eight opens Bill dies. Again, Joan really wants to kill Bill, so she pulls the lever, is lucky enough to open box eight, and Bill dies. Then only the first of these two utterances seems right:

6. Joan intentionally ***killed*** Bill.

6*. Joan intentionally killed ***Bill***.

The focus on ***killed*** in (6) signals that the relevant alternative is [not kill], while the focus on ***Bill*** in (6*) signals that the relevant alternatives are [kill Peter, kill Mary, …]. Pulling the lever increased the probability of Bill dying relative to not dying, but not relative to (e.g.) Peter dying, so the probability-raising account makes the right prediction.

Alternative-sensitivity also has a natural explanation under a causalist account. Linguistic focus is known to affect causal attribution (Shaffer, 2005). Consider:

6'. Bill died because Joan wanted to ***kill*** Bill,

6*'. Bill died because Joan wanted to kill ***Bill.***

The statement seems true only when the focus is on ***kill***. Putting the focus on ***Bill*** signals that the relevant counterfactual alternatives are those where Joan wants to kill someone else. In these counterfactuals, Joan still pulls the lever, and Bill still might die.

Finally we can consider cases where causation and probability-raising dissociate. Suppose that pulling the lever has a 9/10 chance of freeing Bill from his prison cell, and a 1/10 chance of poisoning him. If Joan does nothing Bill will starve to death. But she pulls the lever because she wants to be the one who killed Bill. Pulling the lever triggers the poison release and Bill dies.

Pulling the lever decreased Bill's probability of dying (from 1 to 1/10). Yet:

7. Joan intentionally killed Bill,

7'. Bill died because Joan wanted to kill him.

This result suggests that there can be intentionality without probability-raising. One might of course devise a variant of the probability-raising account that handles such cases, for example by requiring that the agent's action raise the probability relative to a baseline situation where things are 'normal'. But attempts to find a suitable definition of probability-raising might converge to a definition eerily similar to that of causation.

IV) Conclusion

When an agent brings about X as a means to an end Y, people sometimes judge that the agent did Y intentionally but did not do X intentionally. I have suggested that this effect is naturally explained by a causalist theory of intentional action (Quillien & German, 2021). Future research should of course provide more systematic experimental tests of the armchair intuitions I report on the new cases introduced here.

References

Blumberg, K., & Hawthorne, J. (unpublished manuscript) *Kraemer's Puzzle and the Theory of Intentional Action.*

Butler, R. J. (1978). Report on analysis "problem" no. 16. *Analysis, 38(3)*, 113–114.

Davidson, D. (1980). *Essays on Actions and Events: Philosophical Essays Volume 1.* Clarendon Press.

Ericson, S. R., Denison, S., Turri, J., & Friedman, O. (2023). Probability and intentional action. *Cognitive Psychology, 141,* 101551.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological review, 128(5)*, 936.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition, 161*, 80-93.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis, 63(3)*, 190-194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical psychology, 16(2)*, 309-324.

Kraemer, E. R. (1978). Intentional action, chance, and control. *Analysis, 38(3)*, 116–117.

Nadelhoffer, T. (2004). The Butler problem revisited. *Analysis, 64(3)*, 277-284.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology, 61(4)*, 303-332.

Paul, L. A., & Hall, E. J. (2013). *Causation: A user's guide.* Oxford University Press.

Pavese, C., Henne, P., & Beddor, B. (2023). Epistemic Luck, Knowledge-How, and Intentional Action. *Ergo, forthcoming.*

Pavese, C., & Henne, P. (2023). The know-how solution to Kraemer's puzzle. *Cognition, 238*, 105490.

Quillien, T. (2020). When do we think that X caused Y?. *Cognition*

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition, 214*, 104806.

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review.*

Quillien, T., Szollosi, A., Bramley, N. R., & Lucas, C. (2023). Causal inference shapes counterfactual plausibility. *Proceedings of the Cognitive Science Society (45).*

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review, 114(3)*, 327-358.

Planning and Directing

Juan Murillo Vargas and Daniel W. Harris

Mandelkern (2021) notes that utterances like the following are infelicitious.

(1) # Do your homework! But you might not.

This is surprising. There are contexts in which each conjunct is separately thinkable and utter-able, but the conjunction is not. We might imagine a parent directing their child to do their homework while *thinking* that their child might not do it; yet they would still not be able to utter (1).

The data gets weirder. The infelicity remains when deontic modals are used performatively (for an overview see Kaufmann, 2020), when the speaker uses the peremptory mood, and when the conjuncts are flipped.

(2) a. # You must do your homework. But you might not.

b. # I order you to do your homework. But I'm not sure you will.

c. # You won't do your homework. But you must (because I say so).

In contrast, when a deontic modal is used descriptively, or when directive force is weakened (e.g., through a weaker modal, an acquiesive use (von Fintel and Iatridou, 2017), or rising intonation), the infelicity disappears.

(3) a. (It is true that) you must do your homework. But for all I know you won't.

b. You should do your homework. But I'm not sure you will.

c. Do your homework! Don't do your homework! I'm not sure what you'll do.

d. Do your homework? But you might not.

Mandelkern takes these patterns surrounding what we'll call "paradoxical directives" to support a *posturing* norm.

> POSTURING: a speaker ought to issue a directive to an addressee only if she *pretends* to be certain the addressee will comply with the directive.

This explains why a parent can *think* their directive might not be followed, but can't *assert* as much. POSTURING only calls for pretense. And one can pretend to be certain without actually being certain. Further, it correctly predicts that the infelicity remains when the speech-act is a directive—regardless of the order of the conjuncts—and that it disappears when directive force is removed.

We think Mandelkern has uncovered an interesting fact about directive speech-acts. But there's reason to pause over POSTURING. It's a surprising principle, for one: while it's intuitive that directives are governed by *authority* norms, it's less intuitive that they'd be governed by pretense norms. Furthermore, POSTURING doesn't seem fully satisfying. A central goal in linguistics and cognitive science is not just to *describe* our linguistic practices, but to *explain* why they look one way rather than another. POSTURING doesn't quite get us there. While it does a great job in re-describing the puzzling data, it doesn't tell us why directive speech-acts exhibit this behaviour.

POSTURING is thus best seen as a starting point for a fuller explanation of paradoxical directives. Our goal is use independently-motivated principles from the philosophy of action to provide it. We'll start with the assumption that the function of directive speech-acts is to get addressees to form intentions (or plans) (Charlow, 2011, 2014, 2017; Harris, 2022; Roberts, 2023; Grice, 1968). Then we'll consider some additional data Mandelkern doesn't discuss, and use Michael Bratman's theory of joint planning (1992b; 2014) to explain what's going on.

We'll focus on three additional data points. The first is the other speech-acts exhibit similar a pattern. Promises evince this (cf. Ninan, 2005).

(4) a. # I promise to do my homework. But I won't.

b. # I might not do my homework. But I promise to.

c. # I promise to do my homework. But you don't know that!

The second data point is that sometimes we need to issue directives without pretending that they'll be carried out, because we need to convey back-up options. Mandelkern (2021) discusses similar conditionals; but we can see this with directives featuring disjunctions (cf. Starr, 2020; Murray and Starr, 2021) and even without any operators at all.

(5) a. Do your homework. If you don't, you'll fail the class.

b. Do your homework or do the dishes.

c. Don't trespass. Violators will be prosecuted.

The third and final data point is that in certain uncooperative scenarios, directives are issued when the pretense that they'll be followed is immediately undermined. A parent might utter "turn off the TV!" with directive force—only to immediately turn off the TV themselves—without any infelicity.

To explain all the original and new data, we'll use Bratman's theory of joint planning (Bratman, 1992b, 2014, i.a.). Here's a two-paragraph description.

Plans are complex structures of intentions that link our abstract goals to representations of the specific bodily movements by means of which we can accomplish them. In order for plans to serve this function, they must be coherent, both in the sense that the intentions involved are internally consistent with each other, and in the sense that we intend to do things that are consistent with what we believe is possible (or at least what we accept to be possible for planning purposes; see below). This goes for the plans constructed by individual planners. But it also goes for the plans created together by groups of agents.

In the context of joint planning, Bratman argues that groups of agents coordinate their actions by forming shared intentions and then seeking to form "meshing sublans" of those intentions. This is to say that the agents need to have intentions to carry through parts of their shared intention in a way that is intersubjectively coherent. Imagine A and B have a shared intention to make lunch together. Then if A intends to make the salad and expects B to make the soup, B needs to have the converse intention and expectation, or they probably won't coordinate their actions. In this case, we can follow Bratman in saying that A's plan to make salad and B's expectation that A will make salad "support" each other. These support

relations are what constitute meshing subplans, and are an important part of what allows groups of agents to act in coordinated ways.

Suppose, as the authors cited earlier have argued, that directives are used to propose that the addressee adopt an intention, and that these intentions are presupposed by the speaker to be elements in broader shared plans for coordinated action. Then these features of (joint) planning explain both the old and new paradoxical directive data.

First: we can explain why the paradoxical directives in (1) and (2) are infelicitious even when the speaker doesn't believe their directive will be carried out. Bratman (1992a) argues that we are rationally required to make sure our intentions are consistent with what we *accept* in Stalnaker (1984, 2014, i.a.)'s sense. Acceptance in this sense (as we interpret it) is a class of attitudes in which a speaker acts *as if* they believe p for some purpose. While often times we accept p because we believe p, we can also accept p solely for practical reasons, e.g., to be polite (cf. Schiller, 2022). Thus (1) and (2) convey an irrational speaker who issues a plan while also being in an acceptance state that can't support this plan. Hence the infelicity.

Second: we can explain which speech-acts exhibit paradoxicality and which ones don't. Promises are also thought to propose plans (cf e.g., De Kenessey, 2020, i.a.). But not all speech-acts do, including those conveyed in (3). Our account would thus correctly predict that promises pattern with directives whereas other speech-acts don't—as (3) and (4) evince. Any speech-act that proposes a plan in a context that presupposes shared planning is subject to the same rational requirements that explain why (1) and (2) are infelicitous.

Third: since intentions are partial, they often support *back-up plans*: plans to ϕ if some further condition happens to hold, rather than a plan to ϕ *simpliciter*. Such plans are distinct in that they allow that the further condition is possible. (That's part of their functional role and why they're valuable for boundedly rational agents like us!) Our suggestion is that the directives with back-up options in (5) express back-up plans. This explains why they're felicitous unlike their counterparts in (1) and (2).

Finally: Bratmanian joint plans require a background of cooperativity. Both agents must be able to reasonably expect others to act in accordance with their joint plans. But it's a familiar fact of life that not every interlocutor is cooperative. In such circumstances joint plans—and their rational requirements—can't take hold. This explains why directives issued in un-

cooperative situations can be felicitous even when the speaker undermines the expectation they'll be carried out. The rational requirements that would render such directives infelicitous aren't live due to the uncooperative nature of the situation.

References

Bratman, M. (2014). *Shared agency: a planning theory of acting together*. Oxford University Press, New York, NY.

Bratman, M. E. (1992a). Practical Reasoning and Acceptance in a Context. *Mind*, 101(401):1–15. Publisher: [Oxford University Press, Mind Association].

Bratman, M. E. (1992b). Shared Cooperative Activity. *The Philosophical Review*, 101(2):327.

Charlow, N. (2011). *Practical Language: Its Meaning and Use*. University of Michigan.

Charlow, N. (2014). Logic and Semantics for Imperatives. *Journal of Philosophical Logic*, 43(4):617–664.

Charlow, N. (2017). Clause-type, force, and normative judgment in the semantics of imperatives. In Fogal, D., Harris, D., and Moss, M., editors, *New Work on Speech Acts*. Oxford University Press, Oxford.

De Kenessey, B. (2020). Promises as Proposals in Joint Practical Deliberation. *Noûs*, 54(1):204–232.

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. *Foundations of Language*, 4(3):225–242.

Harris, D. W. (2022). Imperative inference and practical rationality. *Philosophical Studies*, 179(4):1065–1090.

Kaufmann, M. (2020). Imperatives. In Gutzmann, D., Matthewson, L., Meier, C., Rullmann, H., and Zimmermann, T., editors, *The Wiley Blackwell Companion to Semantics*, pages 1–42. Wiley, 1 edition.

Mandelkern, M. (2021). Practical Moore Sentences*. *Noûs*, 55(1):39–61.

Murray, S. E. and Starr, W. B. (2021). The structure of communicative acts. *Linguistics and Philosophy*, 44(2):425–474.

Ninan, D. (2005). Two puzzles about deontic necessity. In Nickel, B., Yalcin, S., Gajewski, J., and Hacquard, V., editors, *New Work on Modality*, pages 149–178. MIT Working Papers in Linguistics.

Roberts, C. (2023). Imperatives in dynamic pragmatics. *Semantics and Pragmatics*, 16(7).

Schiller, H. I. (2022). Directing Thought.

Stalnaker, R. (1984). *Inquiry*. MIT Press, Cambridge, Mass.

Stalnaker, R. (2014). *Context*. Context and Content. Oxford University Press, Oxford.

Starr, W. B. (2020). A preference semantics for imperatives. *Semantics and Pragmatics*, 13(6).

von Fintel, K. and Iatridou, S. (2017). A modest proposal for the meaning of imperatives. In Arregui, A., Rivero, M. L., and Solova, A., editors, *Modality Across Syntactic Categories*, volume 1. Oxford University Press.

Contrasting GIVE-causatives in LSF and French

Patricia Cabredo Hofherr (UMR 7023 – *Structures formelles du langage*, CNRS & U. Paris 8)
Adrien Dadone (UMR 7023 – *Structures formelles du langage*, CNRS & U. Paris 8)

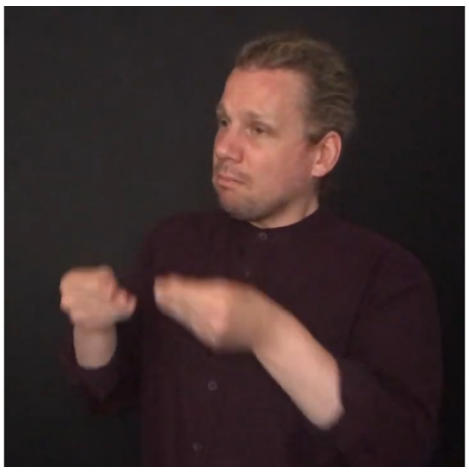
1. Introduction

The present study investigates causative GIVE-constructions in French Sign Language (LSF) in (1a). The causative predicate is lexically identical to the neutral verb of transfer GIVE (1b/c).

(1) a. Pierre **GIVE**$_1$ LAUGH (LSF)
'Pierre makes me laugh.' (causative GIVE)

b. Pierre IX$_a$ BOOK$_{\mathbf{a}}$ $_{\mathbf{a}}$**GIVE**(book)$_1$
'Pierre gives me a book.' (main verb GIVE)[1]

c.




LSF sign GIVE (from Dadone, in progress)

Cross-linguistically, causative constructions using a verb homophonous with the verb of transfer GIVE are common (Lord & al. 2002:223-6, Veenstra & Muysken 2017, see Santoro & Aristodemo 2021 for Italian Sign Language LIS). However, despite the shared lexical origin of the causative verb, causative GIVE constructions (GIVE-CAUSATIVES in what follows) are clearly syntactically and semantically diverse. For example, as illustrated in (2), some GIVE-causatives combine with verb phrases (2a) while others take nouns as complements (2b).

(2) a. Jan **bay** Mari kondwi vwati a (Haitian)
Jean **GIVE** Marie drive car DET
'Jean made Mari drive the car.' (Glaude 2012:170)

b. [Courir comme ça] me **donne** faim. (French)
run like that 1SG.DAT **GIVE** hunger
'Running like that makes me hungry.' (attested)

In what follows, we compare the LSF GIVE-causative and the French GIVE-causative.

[1] Abbreviations in the glosses follow the Leipzig glossing rules with the addition of DET.PTV = partitive determiner. Additional abbreviations in the LSF glosses: IX = pointing sign; a, b: loci in signing space; 1 = signer, 2= interlocutor. $_{\mathbf{a}}$GIVE$_1$ glosses the sign GIVE articulated from locus **a** towards the signer (locus **1**).
LSF does not have tense-marking morphology. We give translations with the simple present in English for sentences that allow past, progressive and future interpretations.

We begin by sketching the language contact situation between LSF and French, motivating the assumption that the LSF GIVE-causative arose as a result of language contact with spoken French (section 2). We proceed to show that the LSF and French GIVE-causatives differ syntactically as well as semantically (sections 3 and 4 respectively). Based on the syntactic properties observed, we propose a syntactic analysis of GIVE-causatives in LSF and French (section 5). Section 6 concludes.

2. Language contact between LSF and French

In general, deaf signers are a linguistic minority in their communities. As a consequence, sign language users are in contact with the spoken and written forms of the languages of their communities (see Millet & Estève 2012 for LSF, Zeshan 2005, Plaza Pust & Morales López eds 2008, Quinto-Pozos & Adam 2020 for general discussion).

In the case of LSF contact with written and spoken French arises in multiple ways. Firstly, in educational contexts, deaf students receive instruction in written French as part of standard schooling. In addition, teaching of French for signers may use a manually coded form of French using signs from LSF (a form of Signed French).[2]

A second source of language contact is bilingual LSF/French speaker-signers, including hearing children of deaf signers and signers that had exposure to spoken French before becoming deaf. Furthermore, many educators in specialised schools for the deaf are hearing signers that learn LSF as part of their training.

Plausibly, LSF GIVE-causatives arose as a result of language contact. Like French GIVE-causatives (3a/b), the LSF GIVE-causative (4) is limited to non-agentive predicates.

(3) a. donner le vertige / donner du souci (French)
GIVE DET vertigo GIVE DET.PTV.SG worry
'to make dizzy / to worry'

b. donner peur
GIVE fear
'to frighten'

(4) a. PIERRE GIVE$_1$ **SAD** (LSF)
'P. makes me sad.'

b. DOG GIVE$_1$ **FEAR**
'Dogs frighten me/ Dogs make me afraid.'

In what follows, we show that despite superficial similarities the GIVE-causative in LSF differs from both French GIVE-causatives syntactically and semantically.

3. Syntactic contrasts between LSF and French GIVE-causatives

French has two GIVE-causatives: one construction with bare *nouns* (5) and one construction with singular or plural NPs denoting emotions (6a) and feelings (6b) (for detailed discussion see Gross 1989).

(5) With bare N: (French)
donner faim / soif / peur / envie
give hunger/ thirst/ fear / desire
'makes hungry / thirsty/ fearful/ make sb want sth'

[2] Signed French is used for educational purposes, reminiscent of linguistic glossing; it is not a natural form of communication between signers.

(6) With NP: (French)

a. **noun expressing an emotion**
donner du souci, du chagrin, des regrets, des remords;
give DET.PTV worry, DET.PTV sadness, DET.INDF.PL regrets, DET.INDF.PL regrets;
un fou rire
an irrepressible laugh
'make worried, sad, regretful / make break out in irrepressible laughter'

b. **noun expressing a feeling**
donner le vertige /la migraine / la nausée / des frissons
give the vertigo / the migraine / the nausea / DET.INDF.PL shivers
'give vertigo / a migraine / nausea / the shivers'
'make dizzy / give a headache / make feel nauseous / make shiver'

French GIVE-causatives have two syntactic forms: the GIVE-causative with a bare N in (5) is a light-verb construction, while the constructions with NPs (6a/b) have the syntax of the lexical verb *donner* 'give'. GIVE-causatives with bare Ns behave as complex predicates, taking the same intensifiers as adjectives (7a) while GIVE-causatives with NPs (7b) take the intensifiers corresponding to NP objects (7c).

(7) a. donner **très** faim / peur (GIVE-causatives + NP) (French)
give very hunger / fear

b. donner **beaucoup de** soucis
give a.lot of worry (GIVE-causatives + NP)

b. donner **beaucoup de** fleurs
give a.lot of flowers (lexical donner 'give' + NP)

GIVE-causatives with NP complements (8) have the same syntax as the lexical verb *donner* 'give' (9):

(8) a. Ce problème **donne des soucis** à Jean. (French)

b. Ce problème lui **donne des soucis**.
this problem (3SG.DAT) gives DET.INDF.PL worries (to Jean).
'This problem worries Jean / worries him/her.' (causative *donner* +NP)

(9) a. Marie **donne des fleurs** à Jean.

b. Marie lui **donne des fleurs**.
Marie (3SG.DAT) gives DET.INDF.PL flowers (to Jean)
'M. gives Jean flowers. / M. gives him/her flowers.' (main verb *donner* 'give')

GIVE-causatives with bare N complements do not allow modification of the noun: with modification a determiner is obligatory (10a), while GIVE-causatives with NP complements admit limited modification (10b).

(10) a. donner soif / donner une soif épouvantable (French)
give thirst / give a thirst terrible
'make thirsty / make terribly thirsty'

b. donner des frissons dans le dos / donner de gros regrets
give DET.INDF.PL shivers in the back give of big regrets
'give shivers down the spine / give great regret'

The syntax of the LSF GIVE-causative differs from both French GIVE-causatives.

Unlike the French GIVE-causative with bare N complements that are always stative, the complements of the LSF construction can be dynamic (11) or stative predicates (12) and may be marked with (non-manual) durative modification (11b).

(11) a. UNHAPPY LOVE STORIES GIVE$_1$ **CRY** (LSF)
'Unhappy love-stories make me cry.'

------durative ---

b. THAT GIVE$_1$ **COUGHING**
'That makes me cough protractedly.'

(12) MEDICATION GIVE **HEALTH GOOD** (attested) (LSF)

While the French GIVE-causative with NP complements follows the syntactic pattern for the lexical verb *donner* 'give' (9a/b), the LSF GIVE-causative differs from the lexical verb GIVE in LSF. The direct complement in the LSF GIVE-causative is obligatorily post-verbal (13a) while for the lexical verb the direct object can be preverbal (13b).[3]

(13) a. PIERRE **GIVE$_1$** **ADVANCE** (LSF)
'Pierre makes me progress.' (GIVE-causative)

b. PIERRE IXa **BOOK** IXa $_a$**GIVE**(book)$_1$
'Pierre gives me a book.' (main verb GIVE)

In contrast with the French GIVE-causative with bare N complements, the LSF GIVE-causative allows modification of the embedded predicate as illustrated in (12) and (14).

(14) TRAINING GIVE$_1$ **CHESS PROGRESS** (LSF)
'The training makes me progress at chess.'

4. Semantic contrasts between LSF and French GIVE-causatives

The LSF GIVE-causative also has a different semantics from the French GIVE-causatives.

While the French GIVE-causatives are limited to emotions (e.g. *peur* 'fear', *soucis* 'worries') and feelings (*migraine* 'migraine', *vertige* 'vertigo'), the LSF GIVE-causative allows a wider range of internally caused changes of state. The LSF GIVE-causative combines with dynamic predicates that can be linked to emotional states (15a/b) but need not be (15c/d/e).

(15) a. SAM GIVE$_1$ **BLUSH** (LSF)
'Sam made me blush.'

b. UNHAPPY LOVE STORIES GIVE$_1$ **CRY/ SAD**

c. ONIONS GIVE$_1$ **CRY /*SAD**
'Unhappy love stories/onions make me cry/sad.'

d. THAT GIVE$_1$ **LEARN STH**
'That makes me learn something' (attested, speaking of a training course)

e. THAT GIVE$_1$ **SNEEZE**
'That made me sneeze.'

Furthermore, the LSF GIVE-causative is compatible with non-psychological internally caused changes of state with predicates like RUST/RUSTY, MELT and DAMAGE (16a/b/c).

(16) a. WATER GIVE METAL **RUSTY** (LSF)
'Water makes metal rust.'

b. SUN GIVE ICE-CUBE **MELT**
'The sun makes ice-cubes melt.'

c. SALT GIVE TREES **DAMAGE**
'Salt causes trees damage.'

[3] The word order with the lexical verb GIVE is flexible: post-verbal objects are possible as in (i). For causative GIVE, in contrast, the order is fixed.

(i) $_1$GIVE$_2$ 400 FRANCS 'I give you 400 francs.' (attested). (LSF)

The locus of the change of state in both French GIVE-causatives in (5/6) is an experiencer of a feeling (*faim* 'hunger') or an emotion (*soucis*, 'worries') and has to be animate. In contrast, the LSF GIVE-causative allows inanimate loci of the change of state (16) in addition to animate experiencers (15).

In an overview of the typology of causative formation, Shibatani (2002:6) identifies 4 classes of verbs, with class 1 most likely to allow causative morphology:

(17) a. CLASS 1 Inactive intransitives (*fall, slip, burn, break, sleep?, laugh*)
b. CLASS 2 Middle/ingestive verbs (*sit down, ascend/ put on clothes, eat, learn*)
c. CLASS 3 Active intransitives (*work, run*)
d. CLASS 4 Transitive (*read the book, paint the house*)

The LSF GIVE-causative does not take complements of classes 2/3/4: Intentional/agentive predicates are ungrammatical.[4]

(18) a. *SAM GIVE KIM **SIT DOWN**. (LSF)
b. *SAM GIVE KIM **RUN**.
c. *SAM GIVE KIM **WASH CLOTHES.**

Only a subset of class 1 predicates is possible in the LSF GIVE-causative. It is not sufficient to have a non-intentional non-agentive predicate: Examples like (19a/b) are not acceptable.

(19) a. *SAM GIVE$_1$ **FALL** (LSF)
Not: 'Sam made me fall' (by startling/pushing me).
b. *SAM GIVE$_1$ **VASE BREAK**.
Not: 'Sam made me break the vase' (by startling/pushing me).

The complements of the LSF GIVE-causative are limited to non-agentive/non-intentional predicates that are construed as internally caused changes of state:

(20) a. psychological predicates (stative HAPPY and non-agentive dynamic: LAUGH)
b. internally caused inactive intransitive (COUGH/SNEEZE/BLUSH*;* LEARN; CRY *(onions)*)
c. internally caused changes of state (RUST/MELT/CRUMBLE) see (16).

The conclusion that the LSF GIVE-causative is limited to predicates of internally caused involuntary change is supported by the contrast found with another means of expressing caused change so-called CLASSIFIER CONSTRUCTIONS in LSF (*TRANSFERTS* in Cuxac 2000, see Garcia & Sallandre 2014 for an overview of the literature on established signs and classifier constructions). Classifier constructions have a strong iconic component. Classifier constructions expressing caused change insist on the process and the result: the process is modified by facial expressions during the time of realization of the predicate. At the same time the higher subject is presented as the cause of the process (21). In contrast, the LSF GIVE-causative expresses only the result, not the process yielding the result: the higher subject is interpreted as a trigger for a change of state, not as a direct cause (22). The examples in (21) and (22) are attested examples from an article in LSF on the effects of the sun:

(21) LSF Classifier constructions expressing caused change

a. SUN LIGHT IX1 RADIATE HORMONE HAPPINESS DEVELOPMENT (LSF)
'Sunlight stimulates the development of the happiness hormone.'

b. EXPOSURE LONGTIME CORNEAL INFLAMMATION
'Continuous [UV-B] exposure causes corneal inflammation.' (attested)
https://www.media-pi.fr/Article/Le-monde-en-LSF/Sante-et-Bien-etre/Les-dangers-du-bronzage/3094

[4] The classes in (17) are formulated as classes of *verbs*. For LSF it is not clear that there is a grammatical distinction between nouns and verbs. The predicates in (18) correspond to the classes in (17) insofar as they are dynamic agentive predicates.

(22) LSF GIVE-causative

a. SUN GIVE$_1$ MOOD POSITIVE (LSF)
'The sun is good for morale.'

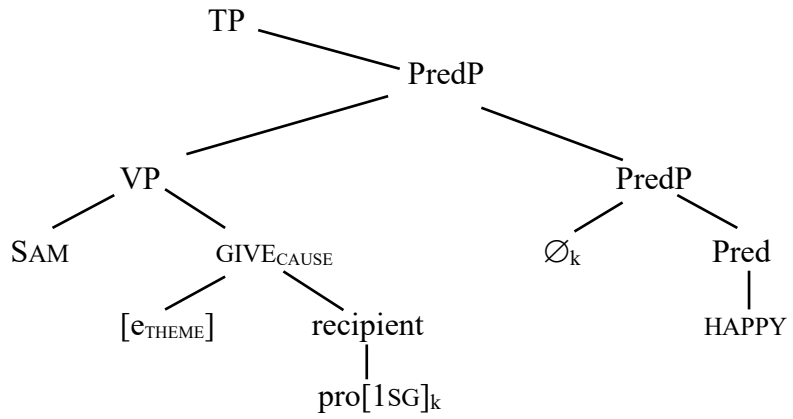
b. ENDORPHINS GIVE$_1$ FEEL GOOD
'Endorphins make us feel good.' (attested)
https://www.media-pi.fr/Article/Le-monde-en-LSF/Sante-et-Bien-etre/Les-dangers-du-bronzage/3094

5. The syntax of GIVE-causatives

LSF GIVE-causatives combine with predicates that allow durative modification and cannot be treated as light verbs.

We propose that LSF GIVE-causatives are adjunction constructions with the GIVE-VP introducing a triggering event related to the causee subject left adjoined to the main phrase denoting a predication of a non-agentive event VP[1-LAUGH] or a result PredP[IX1 HAPPY/IX1 HEALTH GOOD].

(23)



LSF does not have a verb BE/HAVE so the predications [JOHN HAPPY / JOHN GOOD HEALTH] are well-formed as independent phrases in LSF:

(24) SAM HAPPY (LSF)
'Sam is happy.'

SAM HEALTH GOOD
'Sam is in good health/ well.'

The adjunction structure in (23) is compatible with different syntactic forms for the second phrase including dynamic predicates (LAUGH), stative predicates (HAPPY) and nominal predicates (HEALTH GOOD). Adjunction structures like (23) are proposed for serial verb constructions of directed motion in Martinican and Haitian in Zribi-Hertz & al. (2019), following Déchaine (1993).

According to our analysis, the causing relation appears in different syntactic structures in the three GIVE-causatives:

1. in French GIVE-causatives+ NP complement have an abstract transfer with the syntax of the lexical verb *donner* 'give'
2. in French GIVE-causatives+ bare N complements the causative relation is introduced by the causative light verb *give* forming a complex predicate construction
3. in LSF GIVE-causatives a trigger of a change of state is introduced by an adjunction of a VP headed by a grammaticalized GIVE$_{CAUSE}$. The subject of the main predication is not the recipient of the GIVE-verb but an empty category controlled by the recipient.

6. Conclusion

The comparison of French and LSF GIVE-causatives shows that grammaticalization of predicates like GIVE to causative markers is neither syntactically nor semantically uniform.

The French and LSF GIVE-causatives differ in their semantics: while the French GIVE-causatives are limited to emotions and feelings, the LSF construction allows a wider range of predicates of internally caused change including ADVANCE (13a), SNEEZE (15e) and LEARN (15d) with human experiencers but also RUST/ MELT with inanimate loci of change (16).

Acknowledgements We would like to thank the audience of the *Workshop on Serial Verbs Across Modalities* (Paris 8, Sept 2023) and the Seminar *Converging on Causal Ontology Analyses* (Nov. 2023) for comments and suggestions on a previous version of this work.

7. References

Cuxac, Christian 2000. *La Langue des Signes Française (LSF) : les voies de l'iconicité*. Bibliothèque de *Faits de Langues* n°15–16. Paris-Gap: Ophrys.

Dadone, Adrien, in progress. Les subordonnées relatives en « que » et « qui » : analyse d'un point de résistance dans l'accès à la littératie d'adultes sourds locuteurs de la langue des signes française (LSF) langue 1 et proposition didactique de remédiation. PhD in progress. U. Paris 8, Saint-Denis.

Déchaine, Rose-Marie 1993. Predicates across categories. PhD UMass Amherst.

Garcia, Brigitte & Sallandre, Marie-Anne 2014. Reference resolution in French Sign Language. In Cabredo Hofherr, Patricia & Zribi-Hertz, Anne (eds.), *Crosslinguistic studies on Noun Phrase structure and reference,* 316–364. Syntax and semantics, vol. 39. Leiden: Brill.

Glaude, Herby 2012. *Aspects de la syntaxe de l'haïtien*. Paris: Editions Anibwé.

Gross, Gaston 1989. *Les constructions converses du français*. Genève: Droz.

Lord, Carol; Yap, Foong Ha & Iwasaki, Shoichi 2002. Grammaticalization of 'give'. African and Asian perspectives. In Wischer, Ilse & Diewald, Gabriele (eds.), *New Reflections on Grammaticalization*, 217–235. Amsterdam/Philadephia: John Benjamins.

Millet, Agnès & Estève, Isabelle 2012. La querelle séculaire entre l'oralisme et le bilinguisme met-elle la place de la langue des signes française (LSF) en danger dans l'éducation des sourds ? *Cahiers de l'Observatoire des pratiques linguistiques* 3 :161–176.

Plaza-Pust, Carolina & Morales López, Esperanza (eds.) 2008. *Sign Bilingualism: Language Development, Interaction, and Maintenance in Sign Language Contact Situations.* Amsterdam/Philadephia: John Benjamins.

Quinto-Pozos, David & Adam, **Robert 2020.** Language Contact Considering Signed Language. In Grant, Anthony P. (ed.) *The Oxford Handbook of Language Contact*, 679–693. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199945092.013.40

Santoro, Mirko & Aristodemo, Valentina 2021. A preliminary study on causatives in Italian Sign Language. *FEAST* 4: 139–149. **https://doi.org/10.31009/FEAST.i4.11 http://www.raco.cat/index.php/FEAST**

Shibatani, Masayoshi 2002. Introduction: Some basic issues in the grammar of causation. In: Shibatani, M. (ed.) *The grammar of causation and interpersonal manipulation*, 1–22. Amsterdam/Philadephia: John Benjamins.

Veenstra, Tonjes & Muysken, Pieter 2017. Serial verb constructions. In Everaert, Martin & Riemsdijk, Henk van (eds.), *The Wiley-Blackwell companion to syntax*. Oxford: Wiley-Blackwell

Zeshan, Ulrike 2013. Sign Languages. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. (http://wals.info/chapter/s9, Accessed on 2023-12-31.)

Zribi-Hertz, Anne; Jean-Louis, Loïc & Paul, Moles 2019. Left-adjoined bi-valent predicates in two Caribbean French-based creoles: Martinican and Haitian. *Revista Letras* 99:75–100.

Agentivity, Causation and Intention in Passivisation

An empirical case study in Mandarin Chinese and German passive

Ma, Jian
Humboldt-Universität zu Berlin

1 Introduction

Various studies demonstrate that agentivity plays a core role in human cognition and impacts the construction of language structures and processing (e.g. Bornkessel-Schlesewsky & Schlesewsky 2014, Schumacher, Roberts & Järvikivi 2017). However, debates on agentivity and its closely related semantic features are still ongoing. By one influential notion, agentivity is the synergy of various agentive features, e.g. causation, intention etc. (cf. agent prototypicality in Dowty 1991). With this quantitative approach to agentivity, it is assumed that verbs with more agentive features are more compatible in different linguistic constructions (cf. DeLancey 1984, Primus 1999). In previous literature, however, agentivity is widely understood as intention or control over a situation (e.g. Cruse 1973, Verhoeven 2010). This view is more akin to linguistic prominence proposed by Himmelmann & Primus (2015):

(1) *Linguistic Prominence*

a. Linguistic units of equal rank (e.g. semantic features) compete for the status of being in the centre.
b. This status may shift depending on the context.
c. Prominent units function as structural attractors in their domain.

The prominent agentive feature may vary in different linguistic constructions. For instance, causation dominates subject selection in English (Koenig & Davis 2001), whereas in German it is intention (Primus 2012), which also shows a prominent effect in argument alignment (subject-object vs. object-subject) in experiencer object verbs (Verhoeven 2017).

This study turns the lens on passive structures in Mandarin Chinese (Cmn) and German (Deu) and attempts to explore the interactions among the agentive features, i.e. accumulation or prominence. Two parallel acceptability judgment tests in a 7 Likert scale study were designed and implemented in Mandarin Chinese and German, respectively, and aim to disclose (i) whether verbs with different number of agentive features differ according to their quantitative agentivity, and (ii) whether causation or intention plays a prominent role in passivisation.

2 Experiment: Acceptability Judgment Tests

In the experiments, canonical passive forms in target languages were tested, i.e. *bèi*-passive in Mandarin Chinese and *werden*-passive in German, as examples given in (2) and (3). Following Dowty's definition of agentive features, six classes of transitive verbs differing in number of agentive features were constructed: BREAK, WATCH, SEE, HATE, KNOW and EXHIBIT (henceforth B, W, S, H, K and E, the last five classes were adopted from Kretzschmar et al. 2019), see (4). B and W distinguish from each other by [causation], while W can be differentiated from pure sentient verb classes S (perception), H (emotion) and K (cognition) in terms of [intention]. S, H and K are indistinguishable from each other as they all only show the feature [sentience]. Previous studies, however, have shown that they do behave distinctly in different linguistic constructions (e.g. Rapp 1997, Van Valin 1999). They were therefore classified into different verb classes in the present experiments. The last class E (ascription) exhibits no agentive features.

(2) *Mandarin Chinese*
lǐwù bèi dǎkāi.
gift BEI open
'(The) Gift is/was opened.'

(3) *German*
Die Vase wurde zerbrochen.
The vase BECOME broken
'The vase was broken.'

(4) Six tested verb classes with agentive features according to Dowty (1991)

BREAK	WATCH	SEE	HATE	KNOW	EXHIBIT
[causation] [intention] [sentience] [movement]	[intention] [sentience] [movement]	[sentience]	[sentience]	[sentience]	[ϕ]

Based on the research questions, three experimental predictions can be drawn, cf. (5) – (7). ">" indicates that the left verb class is significantly more acceptable than the right, whereas verb classes juxtaposed with "," do not show significant differences in acceptability.

(5) Prediction A: Quantitative Agentivity

BREAK >	WATCH >	SEE, HATE, KNOW >	EXHIBIT
Number of agentive features:			
[4]	[3]	[1]	[0]

(6) Prediction B: Causation

BREAK >	WATCH, SEE, HATE, KNOW, EXHIBIT
Causativity of agent:	
[+causation]	[-causation]

(7) Prediction C: Intention

BREAK, WATCH >	SEE, HATE, KNOW, EXHIBIT
Intentionality of agent:	
[+intention]	[-intention]

In each experiment, 180 critical items, i.e. 6 verb lexemes for each verb class and 5 theme lexemes for each verb lexeme, were constructed following a one-factorial design with six levels for the factor verb class and were distributed evenly in 5 lists. Each list contains 36 critical items and 72 fillers in a randomized order. Word frequency of selected verb lexemes, theme lexemes (NPs) and their co-occurrence were tested prior to the experiment and high word frequency effects were excluded (cf. Jescheniak & Levelt 1994, MacDonald, Pearlmutter & Seidenberg 1994; Word frequency data were retrieved from corpus BCC for Mandarin Chinese and corpus DWDS for German). To avoid and minimise the effects of other sentence elements, only short passives were constructed, i.e. passive without *by*-phrase (no subject of verb in active), and only inanimate (vs. animate) concrete (vs. abstract) NPs were selected.

Both experiments were programmed and performed online on Ibex HU Berlin (`https://korpling.german.hu-berlin.de/ibex/`) and the links to the experiments were sent to native speakers who volunteered to participate in the experiment. Participants were asked to read sentences displayed separately on screen and rate them with a 7 Likert scale for acceptability: 1 for very unacceptable, 7 for very acceptable.

3 Results and Analysis

The distributions of ratings in both experiments (participants: n_{Cmn}=59, n_{Deu}=55, invalid questionnaires excluded) are shown in Figure 1. To analyze the data, I fitted cumulative logit regression model (Christensen 2015) in R, with verb class as fixed effect, participants and verb lexemes as random effects, by-participants random slope for verb class. To obtain an acceptability cline of verb classes, one-to-one pairwise comparisons were conducted among verb classes.

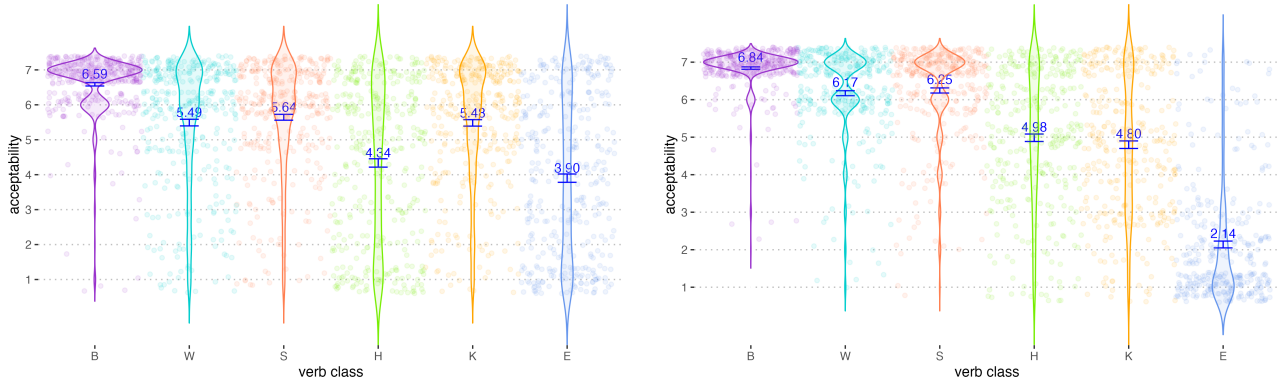


Figure 1: Distribution of ratings among verbs classes with peaks in Mandarin Chinese (left) and German (right) (95% C.I.)

In both languages, B is significantly better accepted than W (Cmn: $p < .05$, Deu: $p < .01$), whereas W and S can not be significantly distinguished from each other (Cmn: $p > .95$, Deu: $p > .85$). Despite the same agentive feature, S, H and K behave quite distinctly and show a cross-linguistic difference, which also appears in E verbs (Cmn: H = E, Deu: H, K > E): In Mandarin Chinese, H is marginally significantly less acceptable than S ($p = .06$) and shows no significance difference with E ($p > .94$). Besides, H is descriptively lower than K, but this difference is not statistically significant ($p > .12$). There is also no significant difference between K and S ($p > .99$) or K and W ($p > .99$). In German, H and K can not be distinguished from each other either ($p > .99$). Both of them are significantly less acceptable than S (S vs. H: $p < .01$, S vs. K: $p < .01$) and more acceptable than E (H vs. E: $p < .01$, K vs. E: $p < .01$). Based on the data analysis, following acceptability clines for each language can be established:

(8) Acceptability cline in *Mandarin Chinese*[1]:

a. B > S, W, K > E

b. B > S, W, > H, E

c. H = K

(9) Acceptability cline in *German*:

B > S, W > H, K > E

Besides the between-group differences, it is also noteworthy that H and E in Mandarin Chinese show within-group differences, which do not appear in German data. Two peaks are shown in the data for H and E verb classes in Mandarin Chinese respectively (cf. Figure 1), which may suggest a two-bias classification within the verb class. This speculation was verified by presenting the acceptability of each verb lexeme in the verb classes H and E, see Figure 2.

Within verb class HATE in Mandarin Chinese, also known as experiencer subject psych verb in the literature (e.g. Landau 2010), verbs expressing the degree of liking/disliking towards objects (e.g. *tǎoyàn* 'dislike') are more acceptable than other emotional verbs (e.g. *hàipà* 'fear'). In the passivisation of verb class EXHIBIT in Mandarin Chinese, the verbs denoting *having* are better accepted than the verbs denoting *lacking*. Note that the latter is expressed in German through intransitive verbs and therefore was not included as experimental items in this study.

[1] By comparing the mean values, an order of acceptability in Mandarin Chinese can be obtained: B, S, W, K, H, E. The between-group differences between S and the subsequent classes W, K are not statistically significant vis-à-vis the difference between S and H, although the difference between H and the adjacent class K is not significant.

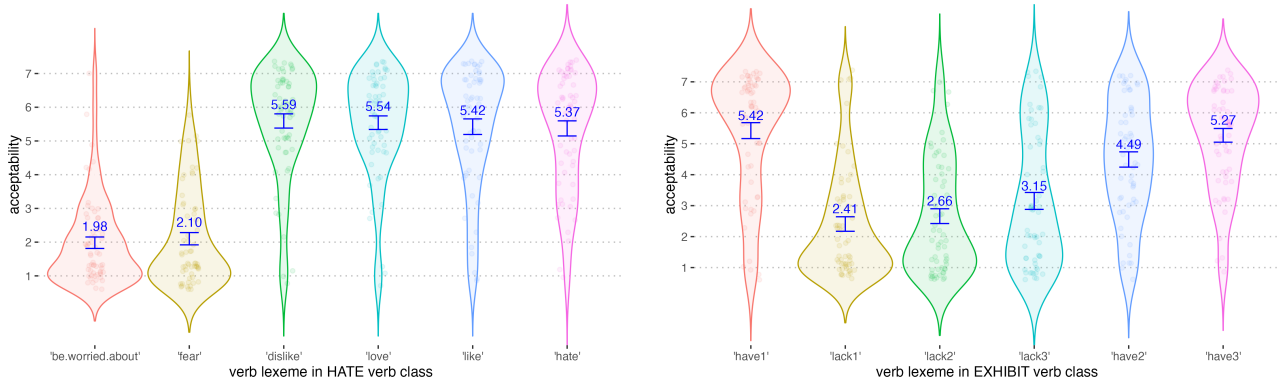


Figure 2: Distribution of ratings within verbs class HATE (left) and EXHIBIT (right) with peaks in Mandarin Chinese (95% C.I.)

4 Discussion

The results of both experiments reject a prediction based on the quantitative agentivity (5) and the effect of intention in passivisation (7). The data seem to suggest a prominent effect of causation (6). The non-causative verbs W and S, however, do show high acceptability (though lower than B). It is hence hardly to draw a conclusion that W and S can not be passivized. One might assume that causation could be gradable and verb classes W and S are causative in a weak sense. Nevertheless, the causation of these "weak causative" verbs contradicts their lack of *change of state*.

Within-group differences in verb classes H and E in Mandarin Chinese also refute the prediction of accumulated agentivity. The presence of two biases within groups suggests that verbs classified in a group based on agentive feature do not always tend to behave consistently, at least not in passive. The reason for this disparity may be that the classification of verbs is not adequately refined, such as state verb class HATE. Some literature argues that these state verbs can be further classified into two kinds of states, Kimian states and Davidsonian states (cf. Maienborn 2005, 2007, 2019, Rothmayr 2009). Davidsonian states show more properties of an action, while Kimian states do not. The two-bias classification appearing in HATE might be due to different state subcategories. In another ongoing study, my colleague and I have found that experiencer subject verbs, i.e. HATE verbs, in Mandarin Chinese and Spanish do have a subset of verbs showing properties of Davidsonian states and another subset showing properties of Kimian states (Ma & Fritz-Huechante in prog.). The reason for the difference within E verb class may be the construction of the verb morphemes. The selected *have* verbs all contain a prepositional verb plus *yǒu* 'have/exist', cf. (10).

(10) Items with verb *-yǒu* '-have/exist' in *Mandarin Chinese*

a. Tǔdì bèi xiǎng-yǒu
land BEI enjoy-have/exist
'(People) have access to land.'

b. Xiànjīn bèi chí-yǒu
cash BEI hold-have/exist
'(People) have cash.'

c. Shǒujī bèi yōng-yǒu
mobile.phone BEI hug-have/exist
'(People) have mobile phones.'

The passivisation of these verbs is more acceptable than *lack*s which may be influenced by the prepositional verb morphemes that can form causative/intentional verbs alone or with other verb morphemes, even though these verbs do not show any of the agentive features defined by Dowty (1991).

In terms of semantic feature dominating passivisation, I suggest that the perspective of analysis should be changed to patient features, since passive forms are patient centered structures. In previous literature, affectedness of the patient is acknowledged as the main factor in passivisation (e.g. Truswell 2008). Despite the close bond with causation, affectedness shows independence and scalarity (e.g. Beavers 2011). The data from both

experiments support the scalar affectedness approach to passivisation from Washio (1993) and Beavers (2011). In order to figure out the effect of affectedness in passivisation, further empirical investigation is needed.

References

Beavers, J. 2011. On affectedness. *Natural language & linguistic theory* 29. 335–370.

Bornkessel-Schlesewsky, I. & M. Schlesewsky. 2014. Competition in argument interpretation: evidence from the neurobiology of language. *Competing motivations in grammar and usage*. 107–126.

Christensen, R. H. B. 2015. Ordinal—regression models for ordinal data. *R package version* 28. 2015.

Cruse, D. A. 1973. Some thoughts on agentivity1. *Journal of linguistics* 9(1). 11–23.

DeLancey, S. 1984. Notes on agentivity and causation. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 8(2). 181–213.

Dowty, D. 1991. Thematic proto-roles and argument selection. *language* 67(3). 547–619.

Himmelmann, N. P. & B. Primus. 2015. Prominence beyond prosody-a first approximation. *pS-prominenceS: Prominences in Linguistics. Proceedings of the International Conference*. 38–58.

Jescheniak, J. D. & W. J. Levelt. 1994. Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *Journal of experimental psychology: learning, Memory, and cognition* 20(4). 824.

Koenig, J.-P. & A. R. Davis. 2001. Sublexical modality and the structure of lexical semantic representations. *Linguistics and Philosophy* 24. 71–124.

Kretzschmar, F. et al. 2019. An experimental investigation of agent prototypicality and agent prominence in german.

Landau, I. 2010. *The locative syntax of experiencers*. London: MIT Press.

Ma, J. & P. Fritz-Huechante. in prog. Passivisation of experiencer subject verbs in Mandarin Chinese and Spanish: The effect of role linkage and state types.

MacDonald, M. C., N. J. Pearlmutter & M. S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review* 101(4). 676.

Maienborn, C. 2005. On the limits of the Davidsonian approach: The case of copula sentences. *Theoretical Linguistics* 31. 275–316.

Maienborn, C. 2007. On Davidsonian and Kimian states. In I. Comorovski & von Heusinger Klaus (eds.), *Existence: Semantics and Syntax*, 107–130. Dordrecht: Springer.

Maienborn, C. 2019. Events and states. In R. Truswell (ed.), *The oxford handbook of event structure, oxford handbooks*, 50–89. Oxford: United Kingdom: Oxford Academic.

Primus, B. 1999. *Cases and thematic roles ergative, accusative and active*. Niemeyer Verlag.

Primus, B. 2012. *Semantische rollen*. Universitätsverl. Winter.

Rapp, I. 1997. *Partizipien und semantische struktur. zu passivischen konstruktionen mit dem 3. status* (Studien zur deutschen Grammatik 54). Tübingen: Stauffenburg Verlag.

Rothmayr, A. 2009. *The structure of stative verbs*, vol. 143. John Benjamins Publishing.

Schumacher, P. B., L. Roberts & J. Järvikivi. 2017. Agentivity drives real-time pronoun resolution: evidence from german er and der. *Lingua* 185. 25–41.

Truswell, R. 2008. Preposition stranding, passivisation, and extraction from adjuncts in germanic. *Linguistic variation yearbook* 8(1). 131–178.

Van Valin, R. D. 1999. Generalized semantic roles and the syntax-semantics interface. *Empirical issues in formal syntax and semantics* 2. 373–389.

Verhoeven, E. 2010. Transitivity in chinese experiencer object verbs. *Transitivity: Form, meaning, acquisition, and processing*. 95–118.

Verhoeven, E. 2017. Scales or features in verb meaning? verb classes as predictors of syntactic behavior. *Belgian Journal of Linguistics* 31(1). 164–193.

Washio, R. 1993. When causatives mean passive: a cross-linguistic perspective. *Journal of East Asian Linguistics* 2(1). 45–90.

Actions, Bodily Movements, and Events in the World

By Robert Reimer[0000-0003-0947-8249]

University of Illinois, Urbana-Champaign

Extended Abstract

Donald Davidson often thought about the right place of causation in action. In his 'Actions, Reasons, and Causes' (Davidson 2001a), he argues that the reasons for an action – mental entities such as beliefs and desires – do not only rationalize but also *cause* that action. He also argues that the kind of causation at play here is event-causality – causality between distinct natural events. It is the kind of causality that we can also find in the context of non-agential phenomena such as collapses of bridges (ibid., 12-13). As the impact of the meteorite caused the collapse of the bridge, so did the pedestrian's desire-event to take a walk together with her belief-event that now it is time to do so cause her action-event of taking a walk.

Davidson's 'Agency' (2001b) can be read as a continuation of the same causalist project. Here, Davidson's interest in the *causes* of actions switches to an interest in actions themselves and their *effects*: "If Brutus murdered Ceasar with the intention of removing a tyrant, then a cause of his action was a desire to remove a tyrant and an effect was the death of Ceasar." (Ibid., 48) Davidson is particularly interested in a sub-class of actions that Joel Feinberg once called '*causally complex act(ion)s*' (Feinberg 2013, 145; see Davidson referring to the same passage 2001b, 56). In performing such an action, the agent physically acts on an object in order to bring about a change or motion in it. Examples for such actions are easily found: Breaking a window or killing someone. Sometimes, such an action comprises a longer causal chain triggered by the agent involving multiple objects. When an agent kills someone, she either physically acts on that person *directly* (for instance by stabbing her) or *indirectly* by acting on some other object directly first (for instance a vial with poison). Anscombe's poisoning gardener is a paradigm example for the latter case (Anscombe 2000, 37).

Again, it is event-causality that connects whatever the agent does with what she thereby brings about: "Causality is central to the concept of agency, but it is ordinary causality between events that is relevant." (Davidson 2001b, 53) According to Davidson, causally complex actions are events sandwiched between the agent's primary reasons and events in the world. But what can such an action be, insofar as it plays this causal role mediating between the agent's mental life and the external world? It must be, just like any other action, some kind of *bodily movement*. Davidson says famously: "We never do more than move our bodies: the rest is up to nature." (Ibid., 59)[1]

Anton Ford (2014, 2018) and Jennifer Hornsby (2011) criticized Davidson's attempt to reduce actions to movements of the body. Ford called Davidson's position '*corporealism*' (Ford 2018, 702) and argued that what I call 'causally complex actions' are more than that. They also *comprise* the events in the object that the agent brings about by physically acting on it. According to Ford and Hornsby, we do not only 'move our bodies' when we act, as Davidson assumes; whatever takes place *out there* in the world is also '*up to us*'.

I agree with Ford and Hornsby's general anti-corporealist position that actions are more than bodily movements. However, it requires a new approach in order to relate actions, bodily movements, and events in the world in a proper way. An essential part of such an approach will be, as I will argue, the drawing

[1] Davidson's position has recently been defended by Adrian Haddock (2005, 162) and Michael Smith (2021).

of *two different* distinction or '*lines*', as I would metaphorically call them. There is a 'vertical line' to be drawn that distinguishes between different causally related events including the agent's bodily movements and events in the world. But there is also a 'horizontal line' to be drawn distinguishing between what the agent *does* (her action) and what thereby *happens* (the causally related events). A causally complex actions is an action that extends into the world *in virtue of* certain causally related events including some bodily movement and some event(s) in the world.

The main purpose of my talk will be to show the necessity of such a 'two-distinction-approach' by revealing certain errors in Davidson's reasoning.

Davidson identifies actions with bodily movements. For him, being causally complex is no *internal* feature of the action itself. That is a peculiar position. Descriptions for such actions ('killing someone' or 'breaking a window') seem to indicate that the caused event (the breakage of the window, the death of the victim) is somehow included in the action. The action of killing someone seems to comprise not only some of the agent's bodily movement but also the death of the victim. But, according to Davidson, the assumption that actions described as killings or breakings comprise more than a bodily movement "springs from a confusion between a feature of the description of an event [the respective action] and a feature of the event itself." (Davidson 2001b, 58)

There is a famous syllogism in Davidson's paper that is supposed to support his corporealist position. Understanding what is wrong with this syllogism is important to see why his position is implausible. In Davidson's text, the queen kills her husband by pouring poison (from a vial) into his ear. Now, Davidson says the following:

(i) "[I]n moving her hand, the queen was doing something that caused the death of the king."
(ii) "Doing something that causes a death is identical with causing a death."
(iii) "But there is no distinction to be made between causing the death of a person and killing him."
(iv) "[T]he killing – took no more time, and did not differ from, the movement of the hand." (Ibid.)

One might want to argue that the claimed identity in (iii) is wrong because the grammatical causatives 'causing a death' or 'causing to die' cannot replace the corresponding lexical causative 'kill' and *vice versa*.[2] This might be correct in *some* cases. However, I think that such replacements work *in general*. Therefore, I grant that the claimed identity in (iii) is correct, in this and in most case.

I am more interested in the following three relations: the relation between (a) the king's death and the queen's action of killing him, (b) the relation between the queen's hand movement and her action of killing him, and, finally, (c) in the relation between her hand movement and the king's death. I think that Davidson does not get any of these relations right. Especially, his position regarding (a) leads to a counterintuitive conclusion.

I will begin with (a). The queen killed the king. She killed him by performing a certain bodily movement. Her killing is *the whole action*. Davidson's piece of reasoning wants to convince his readers that that action is identical to her hand movement. This is what (iv) says. If both (ii) and (iii) are correct, we would have to say that killing someone is an event that caused the victim's death. If we now add the

[2] Byrne did so. He tried to show that lexical causatives 'φ-ing x' (with 'φ' and 'ψ' referring to an action, 'A' referring to an agent, and 'x' referring to an object) cannot be replaced by grammatical causatives such as 'causing x to ψ' because A's φ-ing x is not *always* identical to A's causing x to ψ. Whatever lexical causatives express, it appears to be unique and irreducible (Byrne 2021). Also see J. J. Thomson (1971a, 122) for a similar argument. Indeed, several studies seem to proof these assumption. It seems that speakers of the English language prefer locutions with the grammatical causative 'cause' to the corresponding locutions with lexical causatives if the outcome has been brought about by accident (Wolff 2003) or if the outcome is undesirable or a violation against some law (Sytsma, Bluhm, Willemsen, and Reuter 2019).

further premise that causes *precede* their effects temporally, as Davidson assumes (see Davidson 2001c, 158), we will have to conclude that the queen's action of killing the king took place *before* the king's death. But such a conclusion is counterintuitive.

Many philosophers, including Alvin Goldman (1971, 767), J. J. Thomson (1971), and Paul Pietroski (1998, 77-79), already remarked that the implication of such a temporal difference (between killing and dying) violates the way how we commonly speak and reason. Davidson was well aware of such a potential accusation and tried to downplay it (Davidson 2001d, 177). But I think that the accusation is justified. Where is the mistake that leads to such a strange conclusion?

As I said, I think that (iii) is correct. But I deny that (ii) is correct. As Fred Dretske once said: "Killing a person doesn't cause a person to die. It is a causing, not a cause of death." (Dretske 1988, 37) Let me elaborate a bit on that remark: That killing is a causing of a death means that *in* killing someone, that person is caused to die. So, killing is an action *in the course of which* someone dies. The killing, so to speak, *encompasses* the victim's death. The killing itself, as a whole, is, therefore, *not* the cause of the death. Clearly, since the death is still caused, something else must be its cause. We will come back to the question: "What is the death's cause?" later.

Such an alternative analysis of (a) already solves the problem of temporal order. Since killing is not the cause of the victim's death, it need not be something that is done *before* the death. As something in the course of which someone dies, it could also be something that is done or completed *in the moment* the death has occurred. And that sounds intuitively right to me. Note that this temporal overlap is no accident. It is not *coincidentally* the case that, if someone gets killed, that person dies. There is an important relation of dependence between killing and dying: Without death, there is no killing. Death is a necessary condition for the killing to be successful.

One might wonder whether such an analysis is generalizable. As Rowland Stout remarks, "[p]hilosophy of action has an unhealthy obsession with murder." (Stout 2010, 103) But I think that it is indeed generalizable and does not specifically apply to murder. It is equally true that breakings of windows are causings of breakages of windows but not causes of windows' breakages. What this observation already indicates more generally is the following important point: The relation between actions and events – what an agent *does* (killing someone) and what thereby *happens* (the victim's death) – is *not* a causal relation between temporally consecutive and distinct events. One should rather think of it as a relation of *logical dependence* between entities of two different *ontological categories*. I will say more on what I mean by 'ontological category' later.

Until then, let me turn to (b). When the queen kills the king, she does so by performing something else – a bodily movement. She moves her hand, or to be more precisely, she tilts the vial. How does this movement relate to her action of killing the king? Clearly, both things are things *done* by her. But in contrast to the act of killing, her hand movement is indeed done *before* the king's death. Accordingly, we cannot claim anymore, as Davidson did, that her hand movement and her act of killing are identical to each other. But should we instead say that they are *two* distinct actions?

Such a position has been held by, among others, Arthur Danto. Danto claimed that, when I move a stone with my finger, I perform two actions. First, I perform a bodily movement (which he called 'basic action') – I push the stone. Second, I perform a causally complex action (which he called 'non-basic action') – I move the stone (Danto & Morgenbesser 1963, 463). But I think that this is also counterintuitive. Performing two distinct actions seems to require the onset of two distinguishable bodily movements. Consider a different example: When I bake a cake, I first break eggs and then knead the dough. Here, it makes sense to say that two distinct actions are performed. But it seems weird to say that the person in Danto's example, similarly, first pushes the stone and then moves it.

Let us return to Davidson's example. It seems to be perfectly fine to say that the queen killed the king *by* tilting the vial with her hand. What does the small word 'by' in statements like this express? Many

philosophers, including Goldman (1971, 763), Thomson (1971a, 115; 1971b, 777-778), David Sanford (1984), and Dretske (1988, 38), remarked that it does not express a *causal relation* between two actions. Thomson and Sanford do not even think that it expresses a relation between *two distinct* actions. Sanford, whom I am most sympathetic to, argued that it rather relates two distinct features or characteristics of *one* action. When someone is signaling *by* extending his arm, he does not do two distinct things; he does one thing with two distinct features. His action is an extension of the arm and a signaling (Sanford 1984, 410-411). I think that this is a correct analysis of what 'by' *in general* expresses.[3] But we need to be cautious. Sanford's exemplary action is not causally complex. Here, the by-relation holds between a physical feature of the action – being an extension – and a conventional feature – being a signaling. In the case of Davidson's exemplary action, it must hold between *different kinds* of features.

I think that the by-relation holds between two or more different *stages* of *one* and the *same* action. Such an idea can already be found in the work of von Wright (1971, 68). Recall that causally complex actions are actions that are successfully performed *in virtue of* several events taking place. In the words of Thomson: "[S]omething further has to happen after the shooting in order for the killing to have taken place – *B* has to die." (Thomson 1971a, 131) This is what I already said. Now, if a further event has taken place, the action reaches a new stage. The queen's act of killing, as I said before, is an action that is successfully performed *in virtue of* a death. The death of a person is what makes the queen's action an act of killing. In virtue of the king's death, the queen's action reaches a further stage. It 'becomes' an act of killing.[4]

But the queen's action is *also* an act of tilting a vial because the queen tilts a vial. In fact, the queen kills the king by tilting a vial. Again, we can say that the queen's action is an act of tilting a vial *in virtue of* a certain event taking place. What event is that? I think that it is a certain form of bodily movement –

---

[3] Goldman and Dretske think that 'by' does express a relation between two actions or bits of behavior; but such a relation holds in virtue of a further *causal* relation between one bit of behavior and a causal result of the other bit of behavior. When I ring the bell by pushing a button, the by-relation does hold between my pushing the button and my ringing the bell, but it only holds in virtue of a *causal* relation between my pushing the button and the bell's ringing (Goldman 1971, 763). Thomson, in turn, does not want to speak of *distinct* actions or bits of behavior (Thomson 1971a, 128). Still, she thinks that such a by-statement implies a causal relation between an action and one of its consequences: "Sirhan's shooting of Kennedy certainly caused his death" (ibid., 115). Since I want to deny that causal relations hold between actions and events, as will become clear later, I have to rejects all of these claims here.

[4] I think that the word 'become' is ideal to express the kind of development or 'growth' of actions in stages that I have in mind. Interestingly, I am not the first one choosing such philosophical terminology. Jonathan Bennett also speaks of actions *becoming* other actions in virtue of acquiring new characteristics (Bennett 1973, 316-317). And Hornsby interprets an attempt of Irving Thalberg to explain the causal complexity of the Prime Minister's action of destroying the city of Dauphinia (Thalberg 1977, 110-111) similarly. She writes: "It seems as if he [Thalberg] imagines that the very event that is his moving his finger becomes his destroying the city by the addition to it of other events – just as a tadpole may grow legs and thereby become a thing with legs." (Hornsby 1979, 196)

The analogy between the development of a causally complex action and the metamorphosis of a tadpole might go a little bit too far. In fact, a tadpole does not simply grow additional limbs; it becomes a different animal when it undergoes its transformation. It acquires completely *different* features and not only *more* features. The Prime Minister's action, in turn, does not undergo such a transformation. It retains its original features but 'grows additional limbs' by progresses into a further stage. It becomes more complex.

Hornsby adds that Thalberg must finally reject his assumption that the Prime Minister's moving his fingers *becomes* the destruction of a city because the destruction of a city has parts that the movement of fingers can never have (ibid.) Indeed, Thalberg must draw such a disappointing conclusion because he identifies actions with events. It is clear that the movement of the fingers *as an event* cannot simply stretch out and swallow further bits of events including the collapse and the fire of the city. However, if we understand causally complex actions not as events but as entities in their own rights, entities that develop and 'grow' in virtue of *further* events taking place, Thalberg would not have to drop his assumption. He could say that the Prime Minister's action of moving his fingers *becomes* the destruction of a city *in virtue of* further events taking place – the collapse and the fire of that city.

the queen's hand tilting the vial. This event has to take place in order for the queen's action to be a tilting of a vial. The queen's action first *becomes* a tilting of a vial and then a killing of a person *in virtue of* the respective events taking place. Or in other words: Tilting the vial and killing the king are two different stages of one and the same action in virtue of those two events taking place in the course of that action. The queen's action first reaches the stage of tilting the vial and then the stage of killing the king because her hand tilts the vial (letting the poison drip into his ear), and then the king dies.

The reader might have noticed that I spoke about the queen's bodily movement (the tilting of a vial) ambiguously in the last paragraph. The queen's tilting the vial with her hand is something she *does* and, therefore, a stage of her overall action of killing the king. However, when she does so, something *happens* – her hand tilts the vial. Many philosophers might want to argue that both things are identical. However, such a claim is, at least, not obvious.

I already suggested that, in order to understand causally complex actions properly, we need to draw a general distinction between what agents *do* and what thereby *happens*. The importance of this distinction cannot, in my opinion, be overstated. In contrast to Davidson's distinction, this distinction is not one between distinct and causally related events. It is rather a distinction between entities of different ontological categories. Clearly, those entities take up the same space and occur contemporaneously. Furthermore, there is a relation of logical dependence holding between them. However, it does not follow that those entities are identical to each other.

One might think of this relation in analogy to the one between a university and the university buildings drawn by Gilbert Ryle: The university and all the buildings belonging to it are two (ontologically different) 'things', *despite* taking up the same space (see Ryle 2009, 6). Neither is the university identical to the lecture hall, nor to the library, nor to the seminar building. In fact, it is not even identical to the sum of these buildings.

In the case of the queen's killing the king, there is *one* action performed by the queen logically depending on the occurrence of *several* events – the death of the king, the queen's hand tilting the vial, and possibly more. If we accept the analogy to Ryle's example, we will have to say that the action is not identical to any of them. Indeed, the action consists of several stages *corresponding* to those events taking place. Still, none of these stages is identical to its corresponding event; nor is the action as a whole identical to the sum of all of the corresponding events. Neither is the killing – as the final stage of the queen's action – identical to the king's death, nor is the queen's tilting the vial (with her hand) – as the first stage of her action – identical to her hand's tilting the vial.

This brings me to (c). What is the relation between the queen's hand movement and the king's death? I think that this is indeed the place where *event-causality* comes into play. As I said before, it is true that the king's death was caused by something if he was killed by the queen. What was its cause?

To begin with, the death of the king is a scientifically describable event in the world. Some scientist might be able to give a proper explanation why it occurred by analyzing the causal path of the poison from the vial into his organism. He might even go further and trace back the causal path upstream from the poison in the vial to the queen's contracting muscles, to her nervous system, and all the way up to her brain. It is, in other words, possible to describe the king's death in terms of causal relations between natural events.

Now, it is true that the queen's bodily movement *as* something that *happened* is among these causally related events and, therefore, one of the death's causes. One might be interested in breaking down that bodily movement into further causally related parts – muscle contractions, nerve cell activities, the force of the hand exerted on the vial, etc. Such a breakdown might help the scientist to better embed that movement into the aforementioned causal network of chemical and mechanical events. However, in order to better understand that network, the scientist will not have to look for any actions *among* these

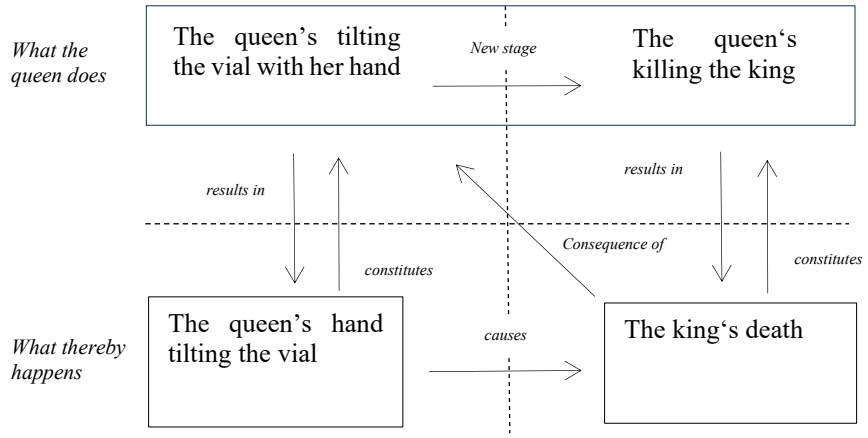
causally related events. In fact, the queen's action *is* not among them. Therefore, as one of the action's *stages*, the queen's tilting the vial – something she did – is not among them, either.

If I may borrow Davidson's wording: "We must conclude, perhaps with a shock of surprise," (Davidson 1971b, 59) that premise (i) is at least misleading. The queen's hand movement – as something that happened – caused the king's death; but the her moving her hand – as a stage of her action – did not.

I reject Davidson's corporealist position that causally complex actions are bodily movements causing further events in the world. Instead, I suggest that actions are entities in their own right logically depending on certain causally related events taking place in- and outside of the agent's body. I am certain that such a position is not only more complex but also more controversial than Davidson's. But I think that it is more plausible.

The key element of my new approach is the distinction between what agents *do* and what thereby *happens*. This is not a causal distinction between two distinct events, as Davidson argues, but a distinction between two entities from different ontological categories. Both entities unfold at the same time and at the same place. Therefore, I will call the distinction between them metaphorically '*horizontal line*'.

Now, there is a further distinction to be drawn on each of these two levels. On the level of what is happening, there is a causal distinction to be drawn between all the *distinct* events in and outside of the agent's body beginning with some of the agent's bodily movements. I will call this distinction metaphorically '*vertical line*'. The unfolding of such a causal chain on the level of what happens is what makes the action causally complex. But such a vertical line must also be drawn on the level of what is done, namely between the different *stages* of the action because these stages are supposed to be expressions of the progress of the underlying causal chain. Here is a visualization of this 'two-distinction-approach':



I think that the general idea of this approach can help to reveal the error in Davidson's corporealist position. Recall Davidson's dictum: "We never do more than move our bodies: the rest is up to nature." If 'up to nature' stands for the natural events that *happen* in the course of an action, as opposed to what we *do*, we can say: Not only what *happens* in the world is up to nature but also what happens in and with our bodies. This is something that our scientist would certainly agree with. I always had a strange feeling when I read Davidson's dictum. As a naturalist, Davidson should say: Of course, it is up to nature what happens in us because we are also part of that nature that we interact with. On the other hand, it is equally true to say that we do not only move our bodies. We also *do* actions that extend *into the world* which means that our performed actions depend on *worldly* events.

The general mistake in Davidson's account, as it is expressed in his dictum, is this: He draws the line between what agents *do* and what is *up to nature* as a 'vertical line' distinguishing different events. But it should be drawn as a 'horizontal line' distinguishing two ontologically different entities – the action as a whole and all the causally related events that it logically depends on.

References

Anscombe, G. E. M. (1957) 2000, *Intention*. Harvard University Press.

Byrne, T. 2021, "Making Metaphysics" *Philosophers' Imprint* 20 (21), 1-18.

Danto, A. C.; Morgenbesser, S. 1963 "What We Can Do." *The Journal of Philosophy* 60 (15), 435-445.

Davidson, D. (1963) 2001a, "Actions, Reasons, and Causes" in *Essays on Actions and Events*, Oxford: Clarendon Press, 3-19.

Davidson, D. (1971) 2001b, "Agency" in *Essays on Actions and Events*, Oxford: Clarendon Press 43-61.

Davidson, D. (1967) 2001c, "Causal Relations" in *Essays on Actions and Events*, Oxford: Clarendon Press, 149-162.

Davidson, D. (1969) 2001d, "The Individuation of Events" in *Essays on Actions and Events*, Oxford: Clarendon Press, 163-180.

Dretske, F. 1988 *Explaining Behavior. Reasons in a World of Causes*. Cambridge: MIT Press.

Feinberg, J. (1964) 2013 "Action and Responsibility" in: Max Black (ed.) *Philosophy in America*. Routledge: New York, 134-160.

Ford, A. 2014, "Action and Passion" *Philosophical Topics* 42 (1), 13-42.

Ford, A. 2018, "The Province of Human Agency" *Noûs* 52 (3), 697-720.

Goldman, A. I. 1971, "The Individuation of Action" *The Journal of Philosophy* 68 (21), 761-774.

Haddock, A. 2005, "At one with our actions, but at two with our bodies: Hornsby's Account of Action" *Philosophical Explorations* 8 (2),157-172.

Hornsby, J. 1979, "Actions and Identities" *Analysis* 39 (4), 195-201.

Hornsby, J. 2011, "Actions in their Circumstances" in Ford, A.; Hornsby, J.; and Stoutland, F. (eds.) *Essays on Anscombe's Intention*, Cambridge, Mass.: Harvard University Press, 105-127.

Pietroski, P. 1998, "Actions, Adjuncts, and Agency" *Mind* 107 (425), 73-111.

Ryle, G. (1949) 2009, *The Concept of Mind*. Oxon: Routledge.

Smith, M. 2021, "Are actions bodily movements?" Philosophical Explorations, 1-14.

Stout, R. 2010, "What Are You Causing in Acting?" In Aguilar, J. and Buckareff, A.: *Causing Human Action: New Perspectives on the Causal Theory of Action*. Cambridge MIT Press, 101-113.

Sytsma, J.; Bluhm, R.; Willemsen, P.; Reuter, K. 2019, „Causal Attributions and Corpus Analysis" In Fischer, E. & Curtis, M (eds.) *Methodological Advances in Experimental Philosophy*. Bloomsbury Press.

Thalberg, I. 1977, *Perception, Emotion and Action*. Blackwell: Oxford.

Thomson, J. J. 1971a, "The Time of Killing" *The Journal of Philosophy* 68 (5), 115-132.

Thomson, J. J. 1971b, "Individuating Actions" *The Journal of Philosophy* 68 (21), 774-781.

Wolff, P. 2003, "Direct causation in the linguistic encoding and individuation of causal events" *Cognition* 88, 1-48.

von Wright, G. H. 1971, *Explanation and Understanding*. London: Routledge & Kegan Paul Ltd.

Demonstratives and the semantic redundancy of intentions
(Jakub Rudnicki)

1. Introduction

The semantics of demonstrative expressions, such as 'this' and 'that', have been in recent decades one of the major points of interest within the debates on context sensitivity. One probable reason for their relevance to these debates is that they might constitute the most natural candidates among all referential expressions for ones whose reference is dependent upon the speaker's intentions.

A major rationale for these suspicions is that, contrary to certain other indexicals, like “I” or “today”, for which it is a knee-jerk reaction to think that their reference is fixed as a function of external features of the context of use (in these cases, perhaps, who is speaking[1] and to what day the moment of the utterance belongs to), there aren’t really any such self-imposing contextual determinants in the case of demonstratives. This intuitive distinction is well captured by John Perry’s (2001: 70) classification into “automatic” indexicals, and “discretionary” ones.

This tendency becomes more pronounced considering that demonstrative pointing, arguably the only potential external feature of context that might automatically determine the semantic values of demonstratives, is notably inadequate for this purpose.[2] There are two main reasons to be skeptical about this idea. Firstly, many seemingly effective uses of demonstratives occur without any accompanying physical demonstrations. Moreover, pointing, as a means of demonstrating, is inherently ambiguous. For instance, when I point at the left sock I am currently wearing, this gesture simultaneously singles out the sock itself, but also its fabric, its color, the Mondrianesque pattern it displays, and even my foot beneath it. Confronted with the natural intuition that users of demonstratives typically and successfully refer to single objects, the discussed view either collapses outright or … necessitates the inclusion of the speaker's intentions to provide disambiguation (Reimer 1992: 387).

In this paper, my focus is not on debating the correctness of viewing the speaker's intentions as crucial for determining the reference of demonstratives. Instead, I will explore an internal debate among those who affirm this view, known as intentionalists about the reference of demonstratives. From the perspective of this paper, intentionalists can be categorized into two distinct groups: *semantic* intentionalists (Stokke 2010; Radulescu 2019) and *pragmatic* intentionalists (Bach 2008; Smit 2012; Heck 2014).

Semantic intentionalism bases its understanding of demonstratives on the model established by standard analyses of automatic indexicals. It posits that the speaker's intention to refer to a specific object is the semantically relevant aspect of the context, which 'automatically' determines the semantic referent of the use of a demonstrative. This approach is classically represented in Kaplan’s (1989b), where he states, 'the referent of a true demonstrative is determined by the utterer’s intention' (p. 585).[3]

Pragmatic intentionalists, in contrast, do not think demonstratives, or their uses in contexts, semantically refer at all. They align more with Strawson's (1950) principle, asserting that it is not the words (in this case, demonstratives) themselves that refer, but rather the speakers who employ them. In other words, on the pragmatic intentionalist approach, the linguistic meaning of demonstratives, or their Kaplanian (1989a) characters, is not sufficiently specific to allow them to determinately refer as a function of context. Instead, demonstrative reference is considered largely a pragmatic phenomenon. This viewpoint is succinctly described by one of its leading proponents as follows:

---

[1] Though see (Ciecierski & Rudnicki 2023) for a recent alternative non-intentionalist view, and (Predelli 1998) for an intentionalist one.

[2] Though, see (McGinn 1981) for a proposal of this sort.

[3] In fact, Kaplan restricts the mentioned analysis to what he calls “perceptual demonstratives”. The view has become an overall intentionalist paradigm in the literature, nevertheless.

> The natural alternative to the claim that demonstratives have meanings that determine their reference as a function of context is that their meanings merely constrain their literal use. Of course, speakers use them to refer, but this can be explained without attributing references to demonstratives themselves. (Bach 2017: 58)

In this paper, I aim to examine a foundational argument that pragmatic intentionalists put forward against the semantic version of intentionalism, which I term the 'semantic redundancy argument'. Having already established the context for this discussion, the structure of the rest of the paper is as follows: Section 2 introduces the redundancy argument against semantic intentionalism. Section 3 presents a seemingly 'obvious' counterexample to this argument in its basic form. In Section 4, I explore the potential responses of pragmatic intentionalists to this counterexample, identifying one as a dead end and another as a promising modification of the redundancy argument that maintains its effectiveness against semantic intentionalism. Finally, Section 5 provides a brief summary of the paper.

2. The "redundancy" argument

The primary concern raised by most pragmatic intentionalists against the semantic intentionalist view is that it inappropriately conflates *pragmatic* or *communicative* phenomena with *semantic* ones. At the heart of their argument is the observation that effective communication fundamentally involves a kind of mind-reading by the hearer. On the classic Gricean (1957) account, for example, the speaker tries to get the hearer to recognize what they are trying to communicate as a result of this very trying. Interpretation, therefore, largely involves discerning the speaker's *intended* message on a specific occasion. This concept of *speaker's meaning* is also applicable to the narrower concept of *speaker's reference*. For instance, if the speaker intends to communicate that 'S is P', the hearer's task of interpretation will be successful only if they manage, among other things, to recognize that the speaker is intending to refer to the object S.

Note that this explanation of communication does not rely on the semantic characteristics of words. While shared semantic meanings undoubtedly facilitate speakers and hearers in their mutual understanding during ordinary linguistic exchanges, the crux of the matter lies in the mind-reading of the speaker's referential and broader communicative intentions. This aspect is fundamentally about the nature of communication itself, rather than the semantics of linguistic expressions.

This perspective, however, clashes with the semantic story of semantic intentionalism. On this view, in the case of utterances involving demonstratives, the interpreter is specifically guided to engage in intention-reading as a matter of these expressions' *semantics*. This is analogous to how the semantics of the word 'I' directs the hearer to identify the speaker of its particular use. However, critics of semantic intentionalism argue that the model's reliance on intention-related instructions is redundant. They contend that such instructions are already inherent in any rational communication and interpretation. J. P. Smit (2012), for instance, posits that a semantic convention, as proposed by semantic intentionalists, contravenes the principle according to which a content of semantic conventions should not involve elements which "serve no communicative or expressive function" (p. 50):

> The problem, simply put, is that, on the Standard view of communication, communication simply is a matter of trying to usefully express, on the part of the speaker, and determine, on the part of the hearer, what the speaker has in mind. Hence, [the hearer] will try to determine the speaker's referent of the uttered demonstrative in virtue of the fact that interpretation just *is* the attempt to determine what a speaker has in mind. The hearer does not need to be told to do so by a linguistic convention. Hence any convention that includes such content violates the [principle]. (p. 52)

A similar sentiment is also expressed by Richard Kimberly Heck (2014):

> It is true that the audience should try to ensure that the referent they assign to an uttered demonstrative is the object to which the speaker intended to refer. But that is not because it is a special fact about *demonstratives* that they always refer to the thing to which the speaker intended to refer. Rather, it is an entirely general principle that, if you want to communicate successfully with the speaker, then you need to ensure that you interpret her words the way she does. This is not only true for other context-dependent expressions […] (p. 344)

3. What's wrong with the "redundancy" argument?

I must say that the mentioned reasoning has always appealed to me as essentially correct. After all, if something looks like pragmatics (or rational communication), and quacks like pragmatics, then it probably simply is pragmatics. Furthermore, given that one needs to posit the mind-reading as involved in communication anyway, insisting on it as supposedly required by the semantics of some expressions, while not others, appears not just redundant but potentially a category mistake (Smit 2012: 54-59).

However, upon further reflection, I have come to realize that the situation may not be as straightforward as it initially appeared, for a quite direct reason. Note that, considering the generality of the observation about the nature of communication, the 'redundancy' argument, at least in its present form, must apply universally to the semantics of all linguistic expressions and their compositional combinations. This line of reasoning, if applicable to the demonstrative 'that', logically extends to all referential expressions and predicates. By the same principle, it should also hold true for complex expressions. This is because the rules governing rational interpretation apply uniformly, whether for simple or complex linguistic forms. The redundancy argument, therefore, must be consistently valid across the entire spectrum of linguistic constructions. In essence, the 'redundancy' argument suggests a universal principle: language should not contain phrases whose reference depends on the speakers' intentions. This is because such a dependency would imply that hearers are being redundantly directed to apply intention-reading, a method they are already employing based on the fundamental principles of rational communication.

This line of reasoning leads to a crucial implication: even a single counterexample could undermine the redundancy argument, at least in its current, unqualified form. Consider the following complex expression, which I believe represents an uncontroversial counterexample:

(1) The object I intend to refer to.

While the expression 'The object I intend to refer to' may not frequently appear in everyday conversations, its also far from artificial. Consider the following story:

> Andy and Betty are enjoying a coffee together. Andy, keen to showcase his knowledge of fancy words and Middle Eastern customs, plans to mention a unique item he saw in a museum. However, as he begins to speak about a particular zarf, the exact word slips his mind. In an attempt to elicit Betty's help and continue the conversation, he says:

(2) Hmm, I forgot the word you know, *the object I intend to refer to* is an ornamental cup-holder.[4]

[4] Although the utterance might sound more natural if Andy referred to 'the object he intends to talk about' instead of 'the object he intends to refer to', readers are free to consider this alternative phrasing. Such a substitution does not affect the underlying argument, as the focus remains on the speaker's intention in determining reference.

I take it to be clear that the reference of the phrase ‘the object I intend to refer to’ in (2) hinges on the specific object that Andy intends to refer to at the time of the utterance. In this scenario, the referent of this complex expression is the zarf he previously saw in the museum.

This finding poses a clear challenge to the general nature of the redundancy argument. Does this imply that the argument is entirely incorrect and the semantic intentionalists can rest assured? As I will explore in Section 4, while the redundancy argument is flawed in its current, unqualified form, I propose that it can be modified. Such a modification would retain its intuitive appeal concerning demonstratives (and other simple expressions), while effectively neutralizing counterexamples like the one presented.

4. A way out of the problem

A natural strategy for someone trying to preserve a version of the ‘redundancy’ argument, albeit in a more restricted form but one still applicable to demonstratives, is to look for differences between demonstratives and complex expressions such “the object I intend to refer to” that could explain why the argument supposedly still applies to the former even if it cannot apply to the latter.

There are (at least) two notable differences to consider. The first pertains to the manner of reference. While both demonstratives and complex expressions like 'the object I intend to refer to' refer in a broad sense (in which the objects of their extensions impact the utterances truth values), only demonstratives engage in 'direct' reference in a narrower sense. In this narrower sense, reference is a semantic relation that directly incorporates the objects referred to into the propositions expressed by the utterances. On the contrary, definite descriptions such as 'the president of the US' are typically not regarded as referring in this narrow, direct sense. Instead, they are understood to 'denote' objects, with their denotation dependent not only on the context of use but also on the world of evaluation. For example, the expression 'the president of the US', as currently used by Joe Biden, denotes Biden in the actual world, but denotes Donald Trump in a possible world where Trump won reelection. In contrast, the directly referential expression 'I', as used by Joe Biden, consistently (directly) refers to Biden in both the actual world and the hypothetical world where Trump won.

The issue with citing the difference between directly referential expressions and denoting expressions as a means to preserve the 'redundancy' argument for demonstratives lies in its apparent irrelevance to the core problem. To claim that the 'redundancy' argument should apply to directly referential expressions but not to denoting ones seems, at least to my understanding, entirely *ad hoc* and unmotivated. There appears to be no clear rationale or compelling justification for this distinction in the context of the argument's validity.

I believe the prospects for successfully refining the 'redundancy' argument are much more promising when considering a second key difference between the two types of expressions. This difference, which I have alluded to earlier, lies in the simplicity of demonstratives as opposed to the semantic complexity of descriptions like 'the object I intend to refer to.' Demonstratives are inherently simple in their semantic structure, while the latter are composed of multiple semantic components.[5] In this context, 'simple' refers to expressions whose semantics are governed by a single convention governing their uses and interpretation. In contrast, 'complex' expressions are those whose semantics emerge from a patchwork of multiple conventions. These conventions, applicable to the individual components of the expression, come together through the application of compositional principles inherent in the language. For instance, to determine the semantic value of a specific use of 'I', it suffices to identify the person who meets the conventional requirement embedded in its linguistic meaning, say, being the speaker of the relevant token. Conversely, the denotation of a complex phrase such as 'the dog' involves synthesizing the meanings of its constituent parts. 'Dog' contributes the set of all dogs, while 'the' imposes a condition of uniqueness.

[5] Well, complex demonstratives, i.e. expressions of the form ‘that F’, are not simple but complex. This does not affect the argument, though, since, if their semantics is assumed to be similar to the semantics of bare demonstratives, it is so due to its being inherited from the latter.

The interplay of these meanings, combined with the context, results in the denotation being a specific, contextually salient dog.

The relevance of this distinction to the 'redundancy' argument lies in the recognition that our languages inherently include ways to talk about the contents of people's intentions. Consequently, it's unavoidable that the reference (in the broader sense) of some phrases will hinge on these intentional contents. This reality, however, does not contradict the claim that no *single* standalone semantic convention of this sort could exist.

In other words, in a sufficiently rich language, in which one can form complex phrases referring to the contents of referential intentions, the "redundancy" argument loses its universal applicability, regardless of its initial plausibility. This, however, does not imply that the fundamental premise of the argument is flawed. My proposal for advocates of the 'redundancy' argument is to reconsider its scope, suggesting that it should primarily apply to simple expressions or semantic conventions as previously defined. In other words, the argument could be almost universally applicable, with the notable exception of expressions where the reference's dependence (in the broader sense) on the speaker's intentions is the outcome of combining separate conventions. Each of these conventions individually adheres to the principle outlined by J. P. Smit.

Alternatively, the fact that the 'redundancy' argument does not extend to some complex expressions can be seen as a circumstantial limitation, arising from the richness and compositional capabilities of language, rather than a fundamental flaw in the argument itself. This perspective suggests that the shortcomings in the argument's applicability are more a reflection of linguistic complexity than an indictment of the argument's core logic.

Or to state it as bluntly as possible, the fundamental insight of the 'redundancy' argument is essentially correct. However, in a language that contains words like 'the', 'object', 'I', 'intend', 'refer', and allows for their combination into a grammatically correct phrase such as 'the object I intend to refer to', the argument simply cannot be universally applicable.

5. Conclusion

In this paper, I pursued two primary objectives. First, I presented a counterexample to the argument proposed by pragmatic intentionalists against the semantic version of intentionalism concerning demonstratives. This argument contends that semantic conventions should not include referential sensitivity to speaker's intentions, as such sensitivity is already inherent in the general rules of rational communication and interpretation. In Section 3, I demonstrated that this argument cannot be universally applied to all types of expressions. I introduced the complex phrase 'the object I intend to refer to' as a clear counterexample, showcasing its inherent referential dependency on the speaker's intention.

In Section 4, I explored potential ways to preserve the core of the 'redundancy' argument and retain its applicability to demonstratives. My proposal was to qualify the scope of the argument, limiting it to simple semantic conventions, and I argued that such a qualification is not arbitrary but justified.

6. References

Bach, K. (2017) Reference, intention, and context. In M. de Ponte & K. Korta (eds.) *Foundations of Meaning*. Oxford University Press. pp. 327–364.

Grice, H. P. (1957) Meaning. *Philosophical Review* 66 (3): 377–388.

Heck, R. K. (2014) Semantics and Context-Dependence: Towards a Strawsonian Account. In B. Sherman & A. Burgess (eds.), *Metasemantics: New Essays on the Foundations of Meaning.* Oxford University Press. pp. 327–364.

Kaplan, D. (1989a) ‘Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and Other Indexicals’, in J. Almog, J. Perry, and H. Wettstein (eds.) Themes from Kaplan, 481–563. Oxford: Oxford University Press.

Kaplan, D. (1989b). Afterthoughts. In J. Almog, J. Perry & H. Wettstein (eds.), *Themes*

From Kaplan. Oxford University Press. pp. 565–614.

McGinn, C. (1981) The mechanism of reference. *Synthese* 49 (2):157–186.

Perry, J. (2001) *Reference and Reflexivity*. Stanford, Calif.: Center for the Study of Language and Inf.

Predelli, S. (1998) ‘I am not here now’, *Analysis*, 58/2: 107–15.

Radulescu, A. (2019) A Defence of Intentionalism about Demonstratives. *Australasian Journal of Philosophy* 97 (4): 775–791.

Reimer, M. (1992) Three views of demonstrative reference. *Synthese* 93 (3):373–402.

Rudnicki, J. (2022) Speaker’s Intentions, Ambiguous Demonstrations, and Relativist Semantics for Demonstratives. *Philosophia* 50 (4):2085-2111.

Rudnicki, J. (2023) Can the reference of a use of “That” change? Assessing nonstandard approaches to the semantics of demonstratives. *Journal of Pragmatics* 209:31-40.

Rudnicki, J. (2023) Semantic conventions and referential intentions. *Synthese* 202 (1):1-16.

Smit, J. P. (2012) Why Bare Demonstratives Need Not Semantically Refer. *Canadian Journal of Philosophy 42 (1):43–66*

Stokke, A. (2010) Intention-sensitive semantics. *Synthese* 175 (3):383–404.

Strawson, P.F. (1950). “On Referring”. *Mind* 59: 320–344.

Are There Speech Acts in Inner Speech?

Daniel Gregory, University of Barcelona

Inner speech only began to receive focused philosophical attention in the last fifteen or so years, but it is now an established subfield in the philosophy of mind. Philosophers have investigated questions about the ontology of inner speech (whether it is really a type of speech or, rather, a mental representation of speech or of some aspect of speech); about its relationship to thought; and about the role which it might play in the acquisition of knowledge of our own mental states, among other issues (see Gregory & Langland-Hassan (2023) for a review of the literature). Another question receiving attention is whether existing insights from the philosophy of language are applicable to inner speech.

One of the major developments in the philosophy of language in the twentieth century was the development of speech act theory. This allowed us to study language-use as involving distinct actions, such as asserting, asking, ordering etc., rather than simply as instances of speaking. So we can ask: Are there speech acts in inner speech? This question divides into three sub-questions:

1. Does inner speech involve actions at all?
2. Can there be speech acts in inner speech given that there is no interlocutor?
3. Is the ontology of inner speech such that it is apt to serve as a medium for speech acts?

I will say something about all three questions, though my emphasis will be on the second. I will also mention a possible complication for the view that I begin to develop by addressing these questions.

Does inner speech involve actions at all?

I have argued previously that inner speech usually does not involve actions (Gregory (2020); cf. Frankfort (2022) and Jorba (forthcoming) for the contrary view). At least as far as the ordinary, apparently unprompted inner monologue goes, inner speech just happens; it is not something we do. I reached this conclusion by arguing that three traditional theories of action are not applicable to inner speech. Ordinarily, we do not produce inner speech for any identifiable reason (see Davidson (1963)); we do not guide our inner speech (see Frankfurt (1978)); and we do not try to produce our inner speech (see O'Shaughnessy (1973) and Hornsby (1980)). Of course, this would not establish that it is impossible that the instances of inner speech which form the ordinary inner monologue are not actions. It may be that these traditional theories of action are not right and that some other theory is. And it could be that the instances of inner speech which form the ordinary inner monologue should be considered actions on the basis of that theory. Nonetheless, I took it that there is very strong reason to doubt that most instances of inner speech are actions.

I will argue, contra my previous position, that a great deal of inner speech consists of actions, albeit actions performed involuntarily. It is widely accepted in the philosophy of action that there can be actions which are performed involuntarily. Paradigm examples are habitual actions and addictive actions. It could reasonably be considered a desideratum for theories of action that they can accommodate involuntary actions of the kinds just mentioned. For these are certainly actions—they are things we do and for which we are responsible—but they are not willed. I previously overlooked the possibility that producing inner speech is also in this category, and was therefore overhasty in drawing this conclusion. Once the possibility that inner speech involves involuntary actions is recognized, it is open to conclude that inner speech involves actions on several theories of action.

Can there be speech acts in inner speech given that there is no interlocutor?

Even supposing that inner speech involves actions, this does not yet establish that there are speech acts in inner speech. That is, it does not yet establish that we do things like make assertions and ask questions in inner speech. It certainly *seems* like we do things like make assertions and ask questions in inner speech, but there is a considerable obstacle to reaching this conclusion. Traditionally, speech act theories presuppose the presence of an interlocutor, and there is obviously no interlocutor in inner speech. It is an open possibility that inner speech involves some kind of action other than speech acts (for example, mental actions). So, does inner speech involve speech acts in particular?

On the classic speech act theory of Searle (1969), speech acts consist in efforts to influence interlocutors in quite specific ways. For example, it is essential to the speech act of making a request that it is an attempt to induce an interlocutor to do something (p. 66). The involvement of an interlocutor is not a formal detail; it is fundamental to the action itself. Among other things, the fact that a request is inherently an effort to induce someone else to do something will inform the particular words a speaker utters and how they say them. For Searle, one cannot make sense of what a request is without reference to an interlocutor.

Wilkinson (2020) agrees that there cannot be speech acts in inner speech understood this way, but argues that there can be speech acts in inner speech on an expressivist speech act theory. Specifically, he cites Strawson (1964) and Bach & Harnish (1979), who understand speech acts in a fashion which is considerably less regimented. They do not analyze speech acts in terms of the particular kinds of effects which an utterance is intended to produce. Rather, they foreground the mental state of the speaker who produces an utterance, treating speech acts as inherently expressions of that mental state.

However, this line of thought confronts the same problem. Although expressivists do foreground speakers' mental states in analyzing speech acts, they nonetheless make critical reference to interlocutors. Bach & Harnish, for example, explicitly hold that expressing involves intending that an interlocutor will recognize that the utterance provides a reason to believe that the speaker is in a particular mental state.[1] This is essential to what the speaker does. So, for expressivists, as with classical speech act theorists, the absence of an interlocutor forecloses on the possibility of performing speech acts in inner speech.

If we wish to vindicate the intuition that we perform speech acts in inner speech, we need to find a plausible theory of speech acts on which speech acts do not critically involve an interlocutor. The speech act theory of Ruth Millikan is such a theory.

Millikan (1998) holds that speech acts are parts of 'conventional' patterns. Conventional patterns are patterns which are reproduced and whose use is maintained by weight of precedent. For a pattern to be reproduced, 'its form [must be] derived from a previous item or items having, in certain respects, the same form' (p. 163). In the context of speech, one convention is that 'when A tells B that p, B responds by believing that p' (Green 2020, Section 4.2). This is the convention in operation when one makes an assertion, and it is reproduced whenever the same kind of transaction happens between two people. A pattern is reproduced by weight of precedent when its previous use explains its current use. What maintains the pattern that one person's assertion (rather than, say, a question) is followed by another's formation of a belief in the proposition asserted is that this is now an established practice. We would not tend to form beliefs in response to others' assertions if there were no such established practice.

Critically, Millikan's theory does not hold that we only perform speech acts because we want to influence others. Linguistic conventions persist because of their iterated reproduction in interpersonal contexts. However, there is no claim that speech acts are only ever performed by someone intending or desiring that someone else will respond in the conventional way. Very often, one will, for example, make an assertion because they intend or desire that someone else will form a belief in response—and they can do this precisely because there is an established convention in place—but this is not an essential part of the speech act. One can perform a speech act neither intending nor desiring to influence a listener; *why* one performs a speech act is a separate question from *whether* one performs a speech act. So, for Millikan, speech acts can be understood without reference to an interlocutor.

This makes room for the possibility that one can perform speech acts in inner speech. One could make an assertion, for example, not because they desire to influence anyone else, but simply because they want to express their mental states (cf. Wilkinson (2020)) or because they anticipate some other benefit to result from doing so.

[1] Thanks to Justin D'Ambrosio for drawing my attention to this.

Is the ontology of inner speech such that it is apt to serve as a medium for speech acts?

Much of the above applies not only to inner speech, but to all private speech, that is, speech produced in the absence of an interlocutor, whether internally or externally. There is no question that external private speech is performed in a medium which is suitable for the performance of speech acts, namely, words spoken aloud. After all, this is the medium in which all external speech is produced. But what about inner speech?

It is a standard tenet of speech act theories that speech acts are not always performed in speech. For example, one can perform the speech acts of greeting someone, of beckoning them to come, or of ordering them to stop what they are doing via appropriate movements of the hand. However, insofar as inner speech consists merely of auditory sensations in the mind, rather than in concrete tokens such as sound waves or hand movements, we might wonder if it is a suitable medium for the performance of speech acts. In fact, insofar as almost all philosophers working on inner speech hold that the auditory sensations in inner speech consist of auditory imagery (the only exceptions I am aware of are O'Brien (2013) and Gauker (2018)), and mental images are usually thought of as representations, one might think that inner speech is an especially poor candidate to serve as a medium for speech acts. Inner speech might be a representation of something which does serve as a medium for speech acts—namely, words spoken aloud—but not such a medium itself. In general, images are not instances of the things they represent.

I will argue that this issue can be overcome again by appealing to Millikan's speech act theory. Once we see that speech acts can be performed without an interlocutor, then any notion that speech acts must involve the production of publicly observable tokens falls away. While we might ordinarily think of mental images as representations of things which occur in the external world, they play a different role in this context.

A Possible Complication: Speech Acts as Reason-Giving

It is very natural to think of speech acts as acts which give reasons. Ordinarily, an assertion gives one a reason to believe something. A question gives one a reason to provide an answer. An order gives on a reason to do something. One might think that, however we think of speech acts, there cannot be speech acts in inner speech, because they are not reason-giving in the appropriate way. I can produce an instance of inner speech which seems like an assertion, a question, or an order, but I do not thereby create any reason to believe something, to answer, or to act which I did not have before.[2]

There are a couple of points to make in response to this. First, although speech acts are often reason-giving, this is not essential to them. If you believe that some proposition is true, and I know that you believe that the proposition is true, and you know that I know that you believe that the proposition is true, then my asserting the proposition does not give you a new reason to believe the proposition. Nor does it give you a new reason to believe that we are aware of one another's mental states in this iterated way, because you already know all of this. So, the fact that instances of inner speech do not provide reasons in the way that many external speech acts give reasons is not a bar to their being speech acts. It is not a *universal* feature of speech acts that they are reason-giving.

Relatedly, I suggest that, even though inner speech acts may not be reason-giving in the way that external speech acts are reason-giving, they nonetheless can have causal influences on our own mental states which are similar to the causal influences that external speech acts can have on the mental states of others. Making an assertion in inner speech may not give me a new reason to believe the proposition that I have asserted, but it may have the effect of making me think about what follows from the proposition, or about the relevance of the proposition in a new context, or about whether I should continue to believe the proposition in the light of recently acquired evidence against it. These are also effects which are produced when one makes an assertion to another individual. They have acquired a reason to believe a proposition, but other things often happen too. There is a good chance, for example, that they will think about what follows from the proposition etc. So, while we may not give ourselves

[2] Thanks to Jean-Moritz Müller for highlighting to me that speech acts—or at least external speech acts—are reason-giving.

reasons when we perform inner speech acts in the same way that we give others reasons when we perform external speech acts, we can nonetheless produce similar kinds of effects.

Conclusion

There is a very strong intuition that we perform speech acts in inner speech. There are a number of possible explanations for this intuition. One is that we do indeed perform speech acts in inner speech. That is, we perform actions in inner speech of the same kind as the actions which we perform in external speech. Another is that we do not perform speech acts in inner speech, but do something similar. A third is that the intuition is quite mistaken: our use of inner speech has very little in common with the performance of speech acts in external speech.

The major obstacle to concluding that there are speech acts in inner speech, and the reason that one might incline towards one of the latter two possibilities, is that speech acts are often thought of as inherently social, and inner speech does not happen in a social context. So, the intuition that we perform speech acts in inner speech will be defensible only if there is some viable theory of speech acts which does not hold that speech acts necessarily happen in a social context. My view is that the speech act theory of Ruth Millikan is such a theory. There is, of course, a much larger question to be answered as to whether Millikan's speech act theory is the best possible theory of speech acts. At the very least, however, it certainly provides a useful framework for thinking further about inner speech and our relationship to it.[3]

References

Bach, K. & R. Harnish (1979), *Linguistic Communication and Speech Acts*, Cambridge, MA: MIT Press.

Davidson, D. (1963), 'Actions, Reasons and Causes', *Journal of Philosophy*, 60: 685-700.

Frankfort, T. (2022), 'Action and Reaction: The Two Voices of Inner Speech', *Teorema*, 41 (1): 51–69.

Frankfurt, H. (1978), 'The Problem of Action', *American Philosophical Quarterly*, 15: 157-162.

Gauker, C. (2018), 'Inner Speech as the Internalization of Outer Speech', in P. Langland-Hassan & A. Vicente, *Inner Speech: New Voices*, Oxford/New York: OUP.

Gregory, D. (2020), 'Are Inner Speech Utterances Actions?', *Teorema*, 39 (3): 55–78.

Gregory, D. & P. Langland-Hassan (2023), 'Inner Speech', in E. N. Zalta & U. Nodelman (eds.), *Stanford Encyclopedia of Philosophy* (Winter 2023 edition), available at https://plato.stanford.edu/entries/inner-speech/.

Green, M. (2020), 'Speech Acts', in E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), available at https://plato.stanford.edu/archives/fall2021/entries/speech-acts/.

Hornsby, J. (1980), *Actions*, London: Routledge and Kegan Paul.

Jorba, M. (forthcoming), 'El Habla Interna en el Marco de las Affordances', in D. P. Chico (ed), *Hacia una Concepción Integral de la Mente: Aportaciones en Filosofia de la Mente y en Ciencia Cognitiva*, Zaragoza: Prensas de la Universidad de Zaragoza.

Millikan, R. G. (1998), 'Language Conventions Made Simple', *Journal of Philosophy*, 95: 161–180.

O'Brien, L. (2013), 'Obsessive Thoughts and Inner Voices', *Philosophical Issues*, 23: 93–108.

[3] I am grateful to participants at the Expressive Speech, Expressive Action workshop which took place at the University of Barcelona in October 2023 for feedback on an earlier version of this paper.

O'Shaughnessy, B. (1973), 'Trying (as the Mental "Pineal Gland")', *Journal of Philosophy*, 70: 365-386.

Searle, J. (1969), *Speech Acts: An Essay in the Philosophy of Language*, Cambridge: CUP.

Strawson, P. (1964/1971), 'Intention and Convention in Speech Acts', in P. Strawson, (1971), *Logico-Linguistic Papers*, reprinted from (1964), *Philosophical Review*, 73: 439-460.

Wilkinson, S. (2020), 'The Agentive Role of Inner Speech in Self-Knowledge', *Teorema*, 39 (2): 7-26.

Conditional Intention Ascriptions

Milan Mossé
University of California, Berkeley

January 2024

Abstract

We introduce a puzzle for conditional intention ascriptions: an unconditional statement is stronger than a conditional one, but statements making unconditional and conditional intention ascriptions often have identical effects on the common ground of a conversation. The same is not true for belief ascriptions. Why then do the conditions of intention ascriptions often come for free, so that one can equally opt for the weaker, conditional ascription or the stronger, unconditional one? It seems that the conditions of intention ascriptions come for free when they are common ground (i.e. commonly accepted) in the context of ascription. But this leads to a second puzzle: if the point of intention ascriptions is to describe an agent's intentions, and their intentions don't depend on what the ascriber accepts, why should ascriptions of the agent's intentions be sensitive to what the ascriber accepts? We develop a semantics for conditional intention ascriptions which resolves both puzzles. On the resulting view, the point of intention ascriptions is to describe an agent's intentions using their expected effects, where ascribers' expectations are informed by the common ground in the context of ascription.

Introduction

These are conditional intention ascriptions (Ferrero, 2015):

(1) The nurse intends to change the patient's dressing, unless he dies.

(2) Ann intends to go to the party this evening, provided Bob goes.

As Ferrero observes, Ann's planning in the afternoon is shaped by her intention, before Bob has decided whether he will go. For example, she may purchase a bottle of wine to bring. Thus (2) does not say that Ann will adopt an intention in the future, if Bob goes.

Further, (2) does not necessarily ascribe an intention with conditional content. Indeed, one can accomplish the ascription (2) by saying

(3) If Bob goes, what Ann intends to do is go to the party as well.

The relative clause in (3) is a "scope island" (May, 1985): it would be strange for "intends" to take scope beyond the relative clause to which it belongs. So it is implausible that (3) attributes to Ann a wide-scope intention [to go to the party if Bob does], and thus implausible that conditional intention ascriptions necessarily ascribe intentions with conditional content.[1]

In sum: the ascription (2) does not say that Ann will later adopt an intention if Bob goes to the party, and it need not say that she now has an unconditional intention with conditional content. It could instead say that she (in some sense) *conditionally intends* to attend the party if Bob does.

First puzzle: free conditions

An unconditional statement is stronger than an conditional one, but the effects on the common ground of conditional and unconditional intention ascriptions are often identical. For example, suppose John intends to buy milk at the store if it is in stock, and that neither he nor I knows whether it is in stock. I can equivalently say either of the following:

(4) John intends to buy milk at the store if it is in stock.

(5) John intends to buy milk at the store.

[1] This is an extension of arguments given by Blumberg & Holguín (2019) for attitudes besides intentions. Ferrero (2009) argues that conditional intentions are not intended conditionals, because the nurse cannot succeed in the intention ascribed by (1) by killing the patient—but this assumes the intended conditional is a material conditional.

The condition of the conditional intention ascription (4) seems to "come for free:" one can leave out the condition and opt for the unconditional ascription (5).

The conditions of conditional belief ascriptions do not similarly come for free. Indeed, suppose John believes that if it rains tomorrow, his daughter Lisa will be angry, because the rain will ruin her chalk art on the sidewalk. Then I can say

(6) John believes that Lisa will be angry if it rains tomorrow.

Suppose you and I know that it will rain tomorrow, but that John does not. I should not say to you

(7) John believes that Lisa will be angry.

In sum: It is puzzling that the conditions of conditional intention ascriptions often come for free, since conditional statements are generally weaker than unconditional ones. After all, this is true for conditional belief ascriptions: their conditions do not come for free.

Second puzzle: sensitivity to ascribers' epistemic states

The conditions of conditional intention ascriptions do not *always* come for free. Suppose Frederic has purchased a lottery ticket and intends to buy a Ferrari if he wins. I can say

(8) Frederic intends to buy a Ferrari if he wins the lottery.

Suppose that, unbeknownst to us, he will win the lottery. I should not say to you

(9) Frederic intends to buy a Ferrari.

Here is another case in which the condition does not come for free. Suppose Sam the Swiftie is looking around the party for Taylor Swift, in order to take a selfie with her, but I have no idea whether she is at the party. It would be misleading for me to say to you, in a matter-of-fact way:

(10) Sam is planning on taking a selfie with Taylor Swift.

This suggests that Swift is at the party. Indeed, it would be natural for you to reply "Wow! I hadn't realized she was here. I'll take a selfie with her as well." It is less misleading for me to say:

(11) Sam is planning on taking a selfie with Taylor Swift if she is here.

The Frederic and Sam cases respectively suggest that conditions do not come for free when they are in fact true, or when they are believed or accepted by the agent. We have seen already in John's case that conditions can come for free when they are not known by the agent or by the ascribers. When then do the conditions of conditional intention ascriptions come for free?

Here is a natural idea that fits with all these cases: conditions come for free when they are common ground in the context of ascription. For example, even if I do not know that milk is in stock, perhaps in the context of ascription we commonly accept that milk is in stock, and thus I can opt for the unconditional ascription (5). By contrast, we do not commonly accept that Frederic will win the lottery or that Swift is at the party, which is why I should not opt for the unconditional ascriptions (9) and (10).

Thus it seems that conditions come for free—that is, that unconditional and conditional ascriptions are both felicitous—when they are common ground (i.e. commonly accepted) in the context of ascription.[2] But this leads to a second puzzle: if the point of intention ascriptions is to describe an agent's intentions,

[2]Throughout, we assume that the common ground is what is commonly accepted in the context of ascription, but the assumption that the common ground is what is commonly accepted rather than (say) what is commonly known will play no role in the arguments that follow.

and their intentions don't depend on what the ascriber accepts, why should ascriptions of the agent's intentions be sensitive to what the ascriber accepts?

Roadmap

The first section of the paper lays out a semantics for conditional intention ascriptions, on which they describe the *conative alternatives* left open by an agent's intentions, given what the agent accepts in the context of their deliberation. Given this semantics, the data regarding Sam the Swiftie are explained pragmatically: it is misleading to say (10), because it falsely implicates that Swift is at the party. We argue against this explanation.

The second section proposes a different semantics for intention ascriptions, on which intentions are modeled not as a set of conative alternatives, but as a set of *hyperplans*, which specify what to do relative to each possible way the world could be. On the proposed view, the point of intention ascriptions is to describe an agent's intentions using their expected effects, where one's expectations are informed by the common ground in the context of ascription. Given this semantics, (10) is misleading, because it makes a false prediction about the effects of Sam's intentions.

1 First model: conative alternatives

We extend Beddor & Goldstein's (2022) semantics for unconditional intention ascriptions to conditional ones (§1.1). We then argue that this semantics cannot be elaborated to explain why intention ascriptions are sensitive to ascribers' epistemic states (§1.2).

1.1 Introducing the conative alternatives semantics

Roughly, this model says that an agent intends that p if q when two conditions are satisfied:

- p is true in the possible worlds in which q is true and the agent's intentions are satisfied; and
- p is an answer to a practical question (e.g. "How to get to work?") facing the agent.

We first define the agent's *conative alternatives*—the worlds in which their intentions are satisfied—and the notion of answering a practical question more formally. We then introduce the semantics.

Definition 1 (Conative alternatives)**.** Fix an agent a and centered world w. Let $\mathsf{Con}_a(w)$ denote the set of a's *conative alternatives* in w; these are the worlds compatible with a's intentions in w. Let $\mathsf{Acc}_a(w)$ denote the set of a's *acceptance alternatives* in w; these are the worlds which a believes or accepts as possible in the context of deliberation in w (Bratman, 1992; Holton, 2014).

We assume that $\mathsf{Con}_a(w) \subseteq \mathsf{Acc}_a(w)$: an agent can intend only what they accept is possible.

Definition 2 (Practical questions)**.** A *practical question* Q is a partition on the set of logically possible worlds. The cells in the partition represent answers to the practical question. For example, "How to get to work?" is a practical question. Its answers are sets of worlds in which the agent gets to work in different ways, for example a set of worlds in which they get there by biking, a set in which they get there by walking, a set in which they get there by taking the bus, and so on.[3]

Definition 3 (Propositional definedness on a question)**.** For an agent a facing a practical question Q in a world w, a proposition p is *defined on* Q (restricted to R) if $[\![p]\!] \cap R$ is a union of cells in $Q|_R$, where $[\![p]\!]$ is the set of worlds in which p is true, and the *restricted question* $Q|_R$ is the set containing each answer to Q intersected with R, whenever that intersection is nonempty.

[3]For simplicity, we model questions as partitions (Hamblin, 1958; Lewis, 1979, 1988), but the discussion here is compatible with other views, for example views on which a question is defined not by an equivalence relation but by a similarity relation, or on which questions are downwards closed collections of sets of possible worlds (cf. Ciardelli, Groenendijk, & Roelofsen 2018).

We extend Beddor & Goldstein's (2022) semantics for unconditional intention ascriptions, by using conditions in conditional intention ascriptions to restrict the conative alternatives (cf. Hintikka 1962):

Definition 4 (Conative alternatives semantics)**.** $[\![a \text{ intends that } p \text{ if } q]\!]^w = 1$ iff:

- $\mathsf{Con}_a(w) \cap [\![q]\!] \subseteq [\![p]\!]$, and
- p is defined on Q, a practical question facing a in w, restricted to $\mathsf{Acc}_a(w)$.[4]

Unconditional intention ascriptions are defined in the natural way:[5]

$$[\![a \text{ intends that } p]\!]^w = 1 \iff [\![a \text{ intends that } p \text{ if } \top]\!]^w = 1$$

This model correctly entails that I can truly say either of the following, when John accepts that the store has milk in stock:

(4) John intends to buy milk at the store if it is in stock.

(5) John intends to buy milk at the store.

More generally, the conative alternatives semantics can be elaborated to provide an answer the first puzzle: conditions come for free when they are accepted by the agent,[6] because the point of intention ascriptions is to describe the agent's intentions from their own point of view, which includes what they accept as true in practical deliberation.

1.2 The problem: sensitivity to ascribers' epistemic states

Now, suppose again that Sam the Swiftie is looking around the party for Taylor Swift, in order to take a selfie with her, but that I have no idea whether she is at the party. It would be misleading for me to say to you, in a matter-of-fact way:

(10) Sam is planning on taking a selfie with Taylor Swift.

This seems to communicate:

(12) Taylor Swift is at the party.

Indeed, it would be felicitous for you to reply to (10) with:

(13) Wow! I hadn't realized she was here. I'll take a selfie with her as well.

How does (10) misleadingly communicate (12), such that (13) is a felicitous reply to it? The simplest explanation is that (10) is false, because it entails (12). But the felicity of the following suggests that (10) does not entail (12):

(14) Sam is planning on taking a selfie with Taylor Swift—but she isn't at the party.

[4] We might add the constraint that q is defined on one of the questions to which a's beliefs are sensitive (Yalcin, 2018). If John does not know that there might be a fire at the store which would prevent him from purchasing milk, it seems I should not say "John intends to buy milk at the store if it isn't on fire."

[5] This is a model of propositional intention-ascriptions: an agent intends *that p*. Some argue that intentions are fundamentally infinitival: any intention is an intention *to* do an action (Baier, 1970, 1977; Mueller, 1979; Ferrero, 2013). As Thompson (2008) observes (fn 44), one can think of the infinitival content of an intention as a set of centered worlds (Lewis, 1979); an infinitive is a centered proposition. So the models considered in this paper involve no loss of generality.

[6] Formally: if q is true throughout $\mathsf{Acc}_a(w)$, then because $\mathsf{Con}_a(w) \subseteq \mathsf{Acc}_a(w)$, it follows that q is true throughout the agent's conative alternatives, and thus $[\![a \text{ intends that } p \text{ if } q]\!]^w = 1$ iff $[\![a \text{ intends that } p \text{ if } \top]\!]^w = 1$.

This does not sound self-contradictory, as it would if (10) entailed (12).

A proponent of the conative alternatives semantics might say that (10) is misleading because it falsely implicates (12), even though (10) is true, because it is true throughout Sam's conative alternatives that he takes a selfie with Swift. This would explain why it it is felicitous to say (14) and why it would not be felicitous to reply to this qualified ascription with (13): implicatures are cancellable, and (12) is not misleading because it cancels the implicature (12) of (10).[7]

However, here is an argument that (10) does not implicate (12). Implicatures are distinguished from presuppositions not only by their cancellability, but by their tendency not to survive various kinds of embedding.[8] But (10) communicates that Swift is at the party when embedded:

(15) Sam isn't planning on taking a selfie with Taylor Swift.

(16) If Sam is planning on taking a selfie with Taylor Swift, I won't try to catch his attention.

(17) Is Sam planning on taking a selfie with Taylor Swift?

(18) Maybe Sam is planning on taking a selfie with Taylor Swift.

(19) Presumably Sam is planning on taking a selfie with Taylor Swift.

Said matter-of-factly, these each suggest that Swift is at the party; it would be felicitous to reply to each with (13). Because implicatures do not survive embedding, this suggests that (10) does not communicate (12) by implicature.[9]

2 Second model: information state-sensitive hyperplans

Now, we propose a semantics for intention ascriptions. An agent's intention state is modeled as a set of *hyperplans*, where a hyperplan maps any information state Acc—representing what is accepted as true in the context of deliberation—to a set of worlds left open by the agent's planning, given that information state. Intuitively, a hyperplan specifies what the agent plans on doing, given what they accept.

On the proposed semantics, I can ascribe to you an intention that p when your hyperplans leave open p as an option, given the information state which is the common ground in the context of ascription. This section will first define hyperplans, the common ground, and the proposed semantics more formally. Then it will address a problem for the proposed semantics.

Definition 5 (Hyperplans)**.** A *hyperplan* $h : (Q, \mathsf{Acc}) \mapsto O$ maps a practical question Q and set of worlds Acc accepted as possible in the context of deliberation to a proposition O, which is defined on Q (restricted to Acc); cf. Yalcin (2019). This proposition O represents the options left open by the hyperplan. Associate each agent a in a centered world w with a set of hyperplans $H_{a,w}$. We will say that an agent a in a world w plans on p, *given* Acc when for all $h \in H_{a,w}$, we have $h(Q, \mathsf{Acc}) \subseteq [\![p]\!]$.[10]

Definition 6 (Common ground)**.** By CG we denote the common ground in the context of ascription. This is a set of possible worlds, which represents what conversants in the context of ascription commonly

[7] A hearer's reasoning from (10) to the implicature (12) might appeal to the relevance of (10) to what Sam will or can do (which might be under discussion); to the fact that intentions are typically successful; or to the fact that if Sam is rational, he will intend only what he believes to be possible. We set the details of this reasoning aside, as the argument that follows does not depend on how they are spelled out.

[8] For discussion of counterexamples, see Recanati (2003).

[9] It might be objected that although intentions may be planning states (Bratman, 1987), the words "intending" and "planning" are used somewhat differently. In particular, saying "Sam is intending to take a selfie with Taylor Swift" does not seem to communicate (12). In reply, we deny the asymmetry between "planning" and "intending": said matter-of-factly, both communicate (12). This is just less clear with "intending" because it is less commonly used in natural language.

[10] Note that a rational agent might plan on p, given $[\![q]\!]$, even though they would not plan on p, were q true. For example, I might plan on going to the top floor of my apartment building, given that a tsunami is coming. But perhaps unbeknownst to me, were a tsunami to come, I would get a text message from the government telling me to run for the hills rather than go on top of my building: I would plan to run for the hills, were a tsunami to come.

accept (i.e. accept, accept that they accept, and so on).

Definition 7 (Information-sensitive hyperplan semantics)**.** We propose that

$$[\![a \text{ intends that } p \text{ if } q]\!]^{w,CG} = 1 \iff \text{for all } h \in H_{a,w}, \text{ we have } h(Q, I \cap [\![q]\!]) \subseteq [\![p]\!],$$

for some practical question Q facing agent a in w, on which p is defined. Here I is an information state, whose default value is CG, the common ground in the context of ascription.

In other words, the ascription that a intends that p if q is true if a plans on p, given q and the common ground. Unconditional intention ascriptions are defined in the natural way:

$$[\![a \text{ intends that } p]\!]^{w,CG} = 1 \iff [\![a \text{ intends that } p \text{ if } \top]\!]^{w,CG} = 1.$$

This semantics is naturally paired with a view on the point of intention ascriptions. On that view, the point of intention ascriptions is to describe an agent's intentions using their expected effects, and one's expectations are informed by the common ground in the context of ascription.[11]

This model can explain why it was misleading for me to say

(10) Sam is planning on taking a selfie with Taylor Swift.

This is true iff Sam plans on taking a selfie with Swift, given what is common ground in the context of ascription. Sam will plan that only if the common ground entails that Swift is at the party. But the common ground does not entail that she is there, since I have no idea whether she is at the party. So the ascription is false: Sam does not plan on taking a selfie with Swift, given the common ground in the context of ascription.[12]

2.1 The problem: agent-centered ascriptions

Recall that this seems fine to say:

(14) Sam is planning on taking a selfie with Taylor Swift—but Swift isn't at the party.

But the above model seems to imply that this utterance is contradictory. To address this problem, we propose that in the first part of this utterance, the relevant information state is shifted from the common ground to Sam's acceptance state. We proceed to explain.

Often, intention ascriptions are made relative to the epistemic states of the agent, and not to those of the ascribers. For example, suppose you and your friend see a child tossing an action figure onto the pavement. You both know that he won't be able destroy it, but his intentions are clear, and you say to your friend:

(20) He intends to destroy the toy.

There need be no suggestion that his plan is compatible with what is common ground in the context of ascription. In such cases, we should see the ascriber as temporarily shifting the context of ascription from the common ground to the acceptance state of the agent. This shift is clearer when you elaborate on your ascription by continuing to speak from the perspective of the child, or by dropping that perspective:

[11] Alternatively, one might say that intention ascriptions are sensitive to the common ground because the point of intention ascriptions is often to provide the agent with advice (cf. Jerzak 2019) or to predict and explain the agent's future behavior. But it is unclear why *ascriptions generally* should be associated with these aims, which can always be accomplished by giving advice or making predictions directly.

[12] Alternatively, one might explain why (10) is misleading by making intention ascriptions sensitive to the acceptance state of the speaker, rather than the common ground (cf. MacFarlane 2014). But suppose it is common ground that I accept that the store has milk in stock and that you do not. Then if I say "John intends to buy milk at the store," it would be odd for you to reply "I don't disagree. It's false that John intends to buy milk at the store." Instead, you might say "I mean sure, he *intends* to buy it at the store." Here it seems you are taking issue with my statement. But if my ascription is relative to my acceptance state and yours is relative to your acceptance state, why should you take issue with my ascription?

(21) He intends to destroy the toy—it's a shame, but it must be done.

(22) He intends to destroy the toy—but he won't be able to.

The utterances (22) and (14) are alike: the first part of the utterance is from the perspective of agent, and the second part drops that perspective. Thus by default I is the common ground CG in the context of ascription, but a speaker can shift I to $\mathsf{Acc}_a(w)$, the acceptance state of the agent who is the subject of the attribution. This can be done by placing emphasis on intends—for example, "Sam *intends* to take a selfie with Swift." A speaker can also shift I to $\mathsf{Acc}_a(w)$ by making an attribution which is evidently contradictory when $I = CG$, and which cannot be cooperatively interpreted as an effort to get the hearer to accommodate the attribution by updating the common ground—this is what happens with (14) and (22).

Shifts away from the default information state can be explained by the claim that the point of intention ascriptions is to describe an agent's intentions using their expected effects, where one's expectations are informed by the common ground in the context of ascription. Indeed, even if p is common ground, one should not expect that the agent will act on their plan, given p, if it is also common ground that the agent will not come to accept that p in the course of pursuing their intention, or that the behavior one is predicting will occur before the agent comes to accept that p. In such cases, one should predict the results of the agent's future actions using the agent's own acceptance state: they will not accept that p in time for them act on their plan, given p.

References

Baier, K. (1970). Act and intent. *Journal of Philosophy*, 67:648–658.

Baier, K. (1977). The intentionality of intention. *Review of Metaphysics*, 30:389–414.

Beddor, B. and Goldstein, S. (2022). A question-sensitive theory of intention.

Blumberg, K. and Holguín, B. (2019). Embedded attitudes. *Journal of Semantics*, 36(3):377–406.

Bratman, M. (1987). *Intention, plans, and practical reason*.

Bratman, M. E. (1992). Practical reasoning and acceptance in a context. *Mind*, 101(401):1–15.

Ciardelli, I., Groenendijk, J., and Roelofsen, F. (2018). *Inquisitive semantics*. Oxford University Press.

Ferrero, L. (2009). Conditional intentions. *Noûs*, 43(4):700–741.

Ferrero, L. (2013). Can I only intend my own actions? In Shoemaker, D., editor, *Oxford Studies in Agency and Responsibility*, volume 1, pages 70–94.

Ferrero, L. (2015). Ludwig on conditional intentions.

Hamblin, C. L. (1958). Questions. *Australasian Journal of Philosophy*, 36(3):159–168.

Hintikka, K. J. J. (1962). Knowledge and belief: An introduction to the logic of the two notions.

Holton, R. (2014). Intention as a model for belief. Oxford University Press.

Jerzak, E. (2019). Two ways to want? *The Journal of Philosophy*, 116(2):65–98.

Lewis, D. (1979). Attitudes de dicto and de se. *The philosophical review*, 88(4):513–543.

Lewis, D. (1988). Relevant implication. *Theoria*, 54(3):161–174.

MacFarlane, J. (2014). *Assessment Sensitivity: Relative Truth and its Applications*. Oxford University Press, Oxford.

May, R. (1985). *Logical form: Its structure and derivation*, volume 12. MIT press.

Mueller, A. (1979). Radical subjectivity. *Ratio*, 19:115–132.

Recanati, F. (2003). Embedded implicatures. *Philosophical Perspectives*, 17(1):299–332.

Thompson, M. (2008). Naïve action theory. *Life and Action*, 2010:85–148.

Yalcin, S. (2018). Belief as question-sensitive. *Philosophy and Phenomenological Research*, 97(1):23–47.

Yalcin, S. (2019). Modeling with hyperplans. *Meaning, Decision, and Norms: Themes from the Work of Allan Gibbard*.

Linguistic Framing Affects Moral Responsibility Assignments Towards AIs and their Creators

Dawson Petersen[1], Amit Almor[1], and Valerie L. Shalin[2]

[1]University of South Carolina, [2]Wright State University

Introduction: Despite the meteoric rise of commercial AI and its growing power in our society, research on human perception of intentionality and responsibility in AI is still lacking. The current study fills this gap by investigating how people assign moral responsibility to AIs using Dennett's (1987) intentional stance approach. Dennett (1987) claims that people understand the behavior of complex systems by reference either to a design stance (i.e., reasoning about its intended function) or an intentional stance (i.e., treating it as a rational agent with beliefs and goals). We tested whether priming participants to adopt either a design or intentional stance towards a language-using AI affected how they assigned moral responsibility to both the AI itself and its creators. This research has applications both for the increasingly important issue of human-AI interaction, and also for basic research questions concerning more general theories of anthropomorphism (Epley et al., 2007; Airenti, 2018).

Literature Review: The capabilities of AI systems has improved rapidly over the last two decades due to increases in computing power, the advent of big data science, and deep learning techniques. AIs in general, and large language models (LLMs) in particular, are very easy to anthropomorphize (i.e., perceive as being human, Mitchell & Krakauer, 2023; Tiku, 2022; Schwitzgebel & Shevlin, 2023). Two factors may be driving this effect: 1) the black box nature of AI and 2) the linguistic abilities of LLMs. As to the first point, normally users' have mechanistic mental models of computer programs which, although flawed, provide causal explanations for the computer's behaviors and predict its outputs (Carroll & Olson, 1988). However, deep learning AIs are not fully understood even by the engineers that build them (Castelvecchi, 2016) because deep learning allows AIs to build their own representations of raw data (LeCun et al., 2015). This makes deep learning AIs a black box in a way that other computer programs are not which can make it difficult for users to build mechanistic mental models and can lead them to switch from a design stance to an intentional stance (Dennett, 1987).

As to the second point, language use seems to be a powerful trigger for the ascription of animacy. Weizenbaum's (1966) primitive ELIZA chatbot showed that people tend to assume that chatbots know much more than they really do and are far more capable than they really are (i.e., the Eliza effect, Hofstadter, 1995). Two main factors may explain this tendency. The first is the uniqueness of human language. Language is a powerful communication system that only humans can use (Hockett, 1959; Hauser et al., 2002). As such, seeing a computer program exhibit apparent linguistic competence may suggest to people that they are dealing with a human. The second factor is pragmatic reasoning. According to most theories of pragmatics (e.g., Grice, 1957; Sperber & Wilson, 1986; Levinson, 2000), hearers must assume that a speaker has a communicative intention in order to interpret their utterance. This basic pragmatic assumption could provide the basis for further elaborative inferences about the speaker's (or AI's) other intentions, beliefs, etc. and thus make it easier to anthropomorphize language AI.

Until now however, very little empirical work directly examines the anthropomorphism of LLMs or deep learning AI more broadly. The majority of empirical work on human-AI interaction focuses on (dis)trust of AI (e.g., Glikson & Woolley, 2020; Troshani et al., 2021; Karataş & Cutright, 2023), and in many cases, researchers fail to clearly distinguish between deep learning AI and more traditional computer programs (e.g., Karataş & Cutright, 2023). As a result, much of the evidence that AI and LLMs are especially likely to be anthropomorphized is intuitive or anecdotal. As such more research is needed to

determine the extent to which contemporary AIs are uniquely easy to anthropomorphize and how this anthropomorphism occurs.

More general theories of anthropomorphism can help illuminate these questions. Early work on anthropomorphism assumed that is was a uniquely childlike error (Piaget, 1926). However, further research has demonstrated that anthropomorphism is an almost universal human tendency among adults across a wide variety of cognitive domains, from perception (Heider and Simmel, 1944; Gao et al., 2010; van Buren et al., 2016), to description (Epley et al., 2007; Epley et al., 2008), to interaction (Airenti, 2018; Zhao and Malle, 2022).

In modern psychology, the dominant theoretical framework for dealing with anthropomorphism draws heavily on Fritz Heider's early work on attribution (1958). Most notably, Heider and Simmel (1944) showed that participants interpret and describe the behavior of simple geometric shapes anthropomorphically when those shapes are animated to act out simple stories. Many more recent studies have replicated this effect (Bassili, 1976; Oatley and Yuill, 1985) and further shown that it occurs when there are temporal contingencies between moving shapes, even if the direction of movement is random. Heider and Simmel (1944) explain this effect in terms of attributions (e.g., causal explanations of behavior). According to attribution theory, people are always attempting to understand why events happen, and attribute the causation of events to various internal (i.e., intentional) and external factors (Kelley & Michela, 1980; Hilton, 2007).

However, this approach is highly fragmented. Specifically, there is a great deal of disagreement about both the underlying mechanism of anthropomorphism and its functionality. Guthrie (1993) and Barrett (2000) argue that anthropomorphism is a cognitive error resulting from the fact that evolutionarily, it is far more costly to fail to notice an agent who is present than to mistakenly attribute agency or intentions where there are none. Airenti (2018) argues than anthropomorphism is a cognitive error resulting from the structural similarity of certain types of interaction to social interaction (i.e., your car failing to start is similar to a human being noncooperative, and therefore results in a social response). In contrast, Epley et al.'s three-factor theory (2007) claims that anthropomorphism is caused by 1) elicited agent knowledge (i.e., resemblance between the entity and a person), 2) effectance (i.e., predictive power), and 3) sociality (i.e., need for human contact). In this view, anthropomorphism is in part an error caused by elicited agent knowledge or sociality, and it is in part a functional strategy for predicting otherwise unpredictable entities.

An alternative to attribution theory stems from Dennett's (1987) book *The Intentional Stance*. Like Epley et al. (2007), Dennett is interested in describing how humans make predictions about the world and argues that humans adopt different predictive strategies depending on the type of system that they are attempting to predict. Very simple systems, such as a ball rolling down a ramp, can be predicted using the physical stance (e.g., naïve physics with its notions of forces and collision is a good predictive strategy for these systems). Other systems, such as computers, are far too complex to be predicted using the physical stance. Instead, humans adopt a design stance. By understanding that a computer is designed to perform certain tasks, one can reason about it in terms that do not reference any of its physical mechanisms and still identify important patterns in its behavior. Dennett proposes that the intentional stance is yet another predictive strategy that allows for reasoning about even more complex systems. To adopt the intentional stance towards any entity is to treat it as a rational agent and ascribe to it beliefs, desires, and goals. Adopting the intentional stance towards an entity does not require one to truly believe that it is conscious, rational, or even that it has intentions, just that reasoning about the entity in intentional terms provides useful predictions about its behavior.

Importantly, in Dennett's theory, many complex entities—such as AIs—may alternatively be conceived of in terms of the design stance or the intentional stance. For example, a user interacting with ChatGPT might ask it for medical advice. If ChatGPT provides a strange or unexpected response, a knowledgeable user adopting the design stance might reason about the goals of the designers (for ChatGPT to produce fluent, contextually relevant English), the techniques involved (prediction of the next word in a sequence, based on patterns in large English corpuses across a variety of genres), etc. Such a user will likely attribute the strange response to an error in the system, and then either disregard the erroneous information (recognizing that ChatGPT is not designed for this use case) or reformat their question in a way that is more likely to generate an accurate answer (i.e., prompt engineering). In contrast, a user adopting the intentional stance would attribute the same response to an intent ("ChatGPT wants to help/harm me") or a knowledge state ("ChatGPT does/doesn't know what it's talking about"). Even if this user correctly identifies that the response provided by the AI is strange, their chain of reasoning could result in a fundamentally different response (such as asking "Are you sure about that?", or concluding that the AI is hopelessly incorrect and never using it again).

In sum, according to many attribution theorists, anthropomorphism is a cognitive error caused by resemblance—either perceptual (Barrett, 2000), conceptual (Epley et al., 2007), or situational (i.e., between interacting with nonfunctioning artifacts and noncooperative people, Airenti, 2018). In contrast, according to the functional accounts—i.e., the intentional stance approach (Dennett, 1987) and Epley et al.'s second factor (2007)—anthropomorphism is an adaptive strategy for predicting complex systems. The functional accounts predict that priming participants to adopt an intentional stance towards an AI should cause them to view the AI as an agent and, therefore, capable of having responsibility for its actions. On the other hand, priming participants to see the AI from a design stance should cause them to view it as a machine and therefore consider its creators to be responsible for its actions. These accounts further predict that participants who are inexperienced with AI should be more likely to view it as an agent since adopting a design stance requires more world knowledge about AI (Dennett, 1987) and because being less familiar with AI makes it less predictable (Epley et al., 2007). While the resemblance error accounts could be compatible with the predicted priming effect, they do not make any prediction with regards to experience because the resemblance between an AI and a person is the same regardless of one's personal experience with AI. Finally, Epley et al. (2007) predicts that individual difference (specifically in the sociality motivation) should cause increased anthropomorphism independently of AI-experience.

The current study tests these predictions using a linguistic framing manipulation. Previous work in the metaphor literature has shown that subtle differences in how information is presented—including grammatical metaphor (i.e., placing a non-agent in an agentive subject position, Devrim, 2015) and voice (i.e., placing an entity as the subject of an active versus passive sentence)—dramatically change how participants evaluate and respond to the situation depicted in a text, even when the propositional content is unchanged. This effect is known as linguistic framing and has been widely reproduced (Thibodeau & Boroditsky, 2011; McGlynn & McGlone, 2019). Notably, participants are typically not aware that they have been influenced by this kind of framing (Thibodeau & Boroditsky, 2013). As such, linguistic framing will provide a valuable test for investigating the anthropomorphism of AI. We expect to find that linguistically framing an AI as an intentional agent will cause people to assign more responsibility to it than when it is framed as a designed system. The functional accounts further predict that this effect will be strongest for participants who have little experience with AI.

Methods: We utilized a judgement priming paradigm in which participants first read a short vignette in one of two linguistic framing conditions and then were asked to make judgements about it. The vignette

(shown in *Table 1*) described how an AI language model "Dr. A.I." gave dangerous health advice causing many patients to be hospitalized and one to die. The linguistic framing manipulation was achieved using grammatical metaphor (i.e., making the AI the grammatical subject of active clauses) as well as active/passive voice shifts. The propositional content of both vignettes was the same. After reading the vignette, participants were asked to rate on a scale from 1-100—1) to what extent the AI, the company that created it, and the patients were each responsible for the outcome, and 2) how much experience they had with language AI. Finally, participants completed the Individual Differences in Anthropomorphism Questionnaire (IDAQ, Waytz et al., 2010), and then were asked to retell the story from the vignette in as much detail as they could remember.

Table 1. Intentional and Design Condition Vignettes

Intentional Condition	Design Condition
In 2023, an A.I. language model called "Dr. A.I." captured widespread attention after being released by a tech company called Health A.I. Dr. A.I. tried to provide accurate, tailored medical advice based on what it knew about users' symptoms and medical histories. However, in 2024, Dr. A.I. made an error when it recommended a dangerous home cure for a common cold. Several people who followed this advice were hospitalized, and one person died. The families of the people who were hospitalized are preparing a large lawsuit against Health A.I.	In 2023, a tech company called Health A.I. captured widespread attention after they created an A.I. language model called "Dr. A.I.". Dr. A.I. was designed to provide accurate, tailored medical advice based on the company's data about users' symptoms and medical histories. However, in 2024, a recommendation for a dangerous home cure for a common cold was generated by Dr A.I. Several people who followed this advice were hospitalized, and one person died. The families of the people who were hospitalized are preparing a large lawsuit against Health A.I.

Table 1. Table 1 shows the vignettes for both conditions. Key differences between them are underlined.

Results: We recruited 157 participants from psychology and linguistics classes at the University of South Carolina. Of these, 35 were excluded for failure to complete the study or failure to recall the key details of the vignette, resulting in a final sample size of 122. The data were analyzed in R 4.3.0 (R Core Team, 2023). Overall, participants assigned the most responsibility to the company (M = 70, SD = 23), followed by the AI (M = 49, SD = 35), and least to the patients (M = 43, SD = 26) (illustrated in *Figure 1*).

Figure 1. Mean responsibility assignments by target and AI Experience.

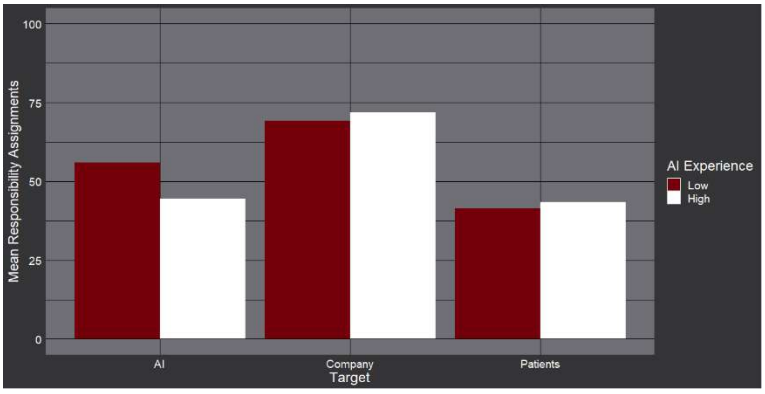
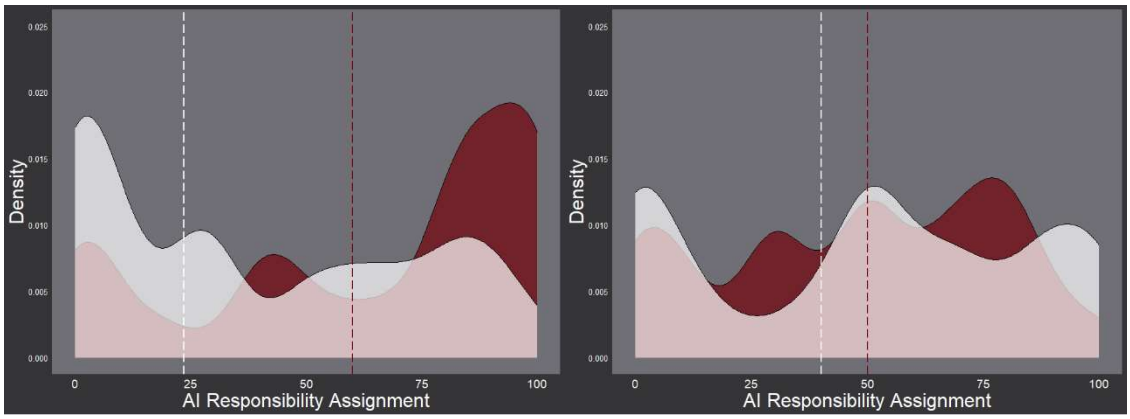


Figure 1. Figure 1 shows the mean responsibility (y-axis) rated on a scale from 1-100 assigned to each target.

Because the rating data were not normally distributed, we analyzed them with cumulative link regression models (Agresti, 2012) using the *ordinal* package (Christensen, 2022). Each dependent variable (responsibility assigned to the AI, the company, and the patients) was modeled using condition (intentional vs design) and log self-rated language AI experience as predictors. Participants' IDAQ scores were not included as they failed to improve the fit of the models. For AI responsibility, we found a main effect of AI-experience ($z = -3.68$, $p < .001$) such that participants with less AI-experience assigned more responsibility to the AI and an interaction between condition and AI-experience ($z = 2.13$, $p = .032$) such that low AI-experience participants assigned more responsibility to the AI in the intentional condition than the design condition, while high AI-experience participants did not (illustrated in *Figure 2*).

Figure 2. Responsibility Assigned to the AI in the Intentional and Design Conditions

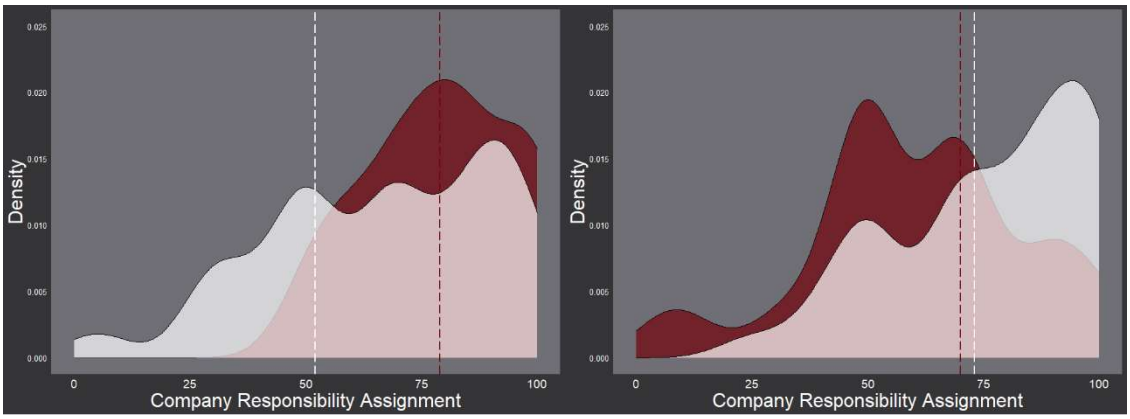


Intentional Condition Design Condition

Figure 2. On the x-axis, Figure 2 shows the distribution of AI responsibility assignments for low (red) and high AI-experience participants (white) with density on the y-axis. Medians are shown by the dashed lines. The left graph shows the results in the *Intentional Condition*, and the right graph shows the results in the *Design Condition*. Low AI-experience participants rated the AI as *more* responsible in the intentional condition than the design condition, while high AI-experience participants did not.

For company responsibility, we found a main effect of condition ($z = -2.01$, $p = .036$) such that participants in the intentional condition assigned less responsibility to the company, and an interaction between condition and AI-experience ($z = 2.42$, $p = .015$) such that the main effect of condition was stronger for participants with high AI-experience (illustrated in *Figure 3*). We found no effects on patient responsibility (illustrated in *Figure 4*).

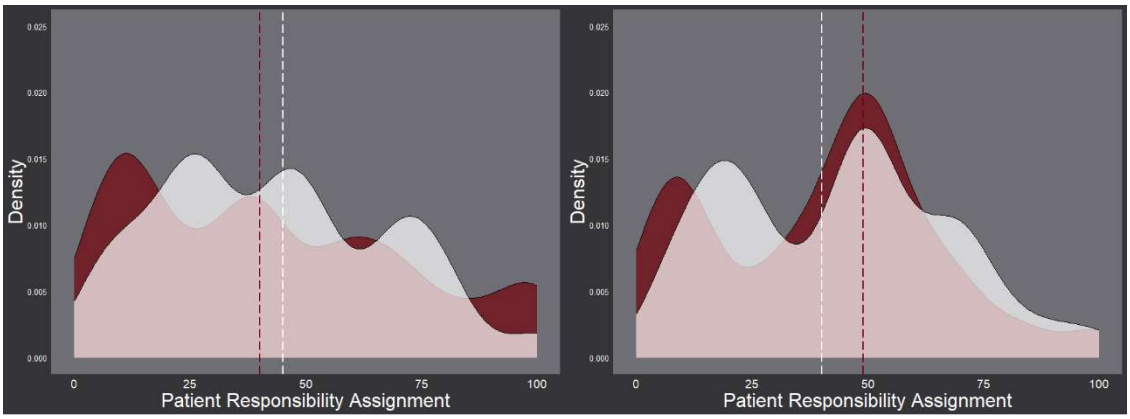
Figure 3. Responsibility Assigned to the Company in the Intentional and Design Conditions



Intentional Condition Design Condition

Figure 3. On the x-axis, Figure 3 shows the distribution of company responsibility assignments for low (red) and high AI-experience participants (white) with density on the y-axis. Medians are shown by the dashed lines. The left graph shows the results in the *Intentional Condition*, and the right graph shows the results in the *Design Condition*. High AI-experience participants rated the company as *less* responsible in the intentional condition than the design condition, while low AI-experience participants did not.

Figure 4. Responsibility Assigned to the Patients in the Intentional and Design Conditions



Intentional Condition Design Condition

Figure 4. On the x-axis, Figure 4 shows the distribution of patient responsibility assignments for low (red) and high AI-experience participants (white) with density on the y-axis. Medians are shown by the dashed lines. The left graph shows the results in the *Intentional Condition*, and the right graph shows the results in the *Design Condition*. No significant effects were found on patient responsibility assignments.

Discussion: Overall, our findings are most consistent with the functional accounts of anthropomorphism found in Dennett (1987) and Epley et al. (2007). Participants with less AI-experience were more likely to anthropomorphize the AI by assigning higher responsibility to it. Furthermore, this between groups difference increased in the intentional condition, showing that low experience participants, but not high experience participants, were quick to adopt an intentional stance towards the AI when primed to do so using linguistic framing. These findings are inconsistent with the cognitive error accounts found in Barrett (2000) and Airenti (2018). A multifactor theory, such as that as Epley et al. (2007), may still be correct.

However, Epley et al.'s prediction that were would be individual differences in anthropomorphism based on the IDAQ was not born out as it showed no significant effects on responsibility assignments. Therefore, our findings are most consistent with the purely functional account of Dennett (1987).

Additionally, we found a result we did not expect—namely that although high AI-experience participants did not assign more responsibility to the AI as a result of our manipulation, they did assign *less* responsibility to the company in the intentional condition than in the design condition. Although unexpected, this finding is in some ways consistent with Dennett's account if it is the design stance priming which caused these participants to assign *more* responsibility to the creators because the design stance highlights the role of the designer. However, in Dennett's account, the stances are meant to be categorical. Therefore, it is difficult for Dennett to explain what the high experience participants were doing in the intentional condition as they assigned low responsibility to both the AI and its creators.

Our findings also have important implications for human-AI interaction. Firstly, we found that anthropomorphism of AI was high overall, especially for low experience participants. While participants assigned the most responsibility to the company, only 15% of participants assigned no responsibility at all to the AI, and on average participants assigned more responsibility to the AI than to the patients who took its advice. This is consistent with the idea that AIs are easy to anthropomorphize. However, further research is needed to compare the anthropomorphism of LLMs to other AI and non-AI programs. Until then, we cannot say to what extent the black box nature of AI and the use of language each contribute to this anthropomorphism. Finally, our unexpected finding—that experienced participants assign low responsibility to the AI's creator when primed to anthropomorphize it—is potentially quite troubling. Historically, authors disagree as to the extent to which such anthropomorphism of AI is desirable (Deshpande, 2023) or dangerous (Hasan, 2023). Indeed, some AI researchers even advocate including anthropomorphic features to increase user trust in the AI (Song & Luximon, 2020). Given our findings, this is a dangerous trend as it could cause even experienced individuals to fail to hold AI companies accountable when their creations cause harm.

References

Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02136

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Religion and Cognition: A Reader*, *4*(1), 86–98.

Bassili, J. N. (1976). Temporal and spatial contingencies in the perception of social events. *Journal of Personality and Social Psychology, 33*(6), 680–685. https://doi.org/10.1037/0022-3514.33.6.680

Carroll, J. M., & Olson, J. R. (1988). Mental models in human-computer interaction. *Handbook of human-computer interaction*, 45-65.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News, 538*(7623), 20.

Dennett, D. C. (1987). *The Intentional Stance*. MIT.

Deshpande, A., Rajpurohit, T., Narasimhan, K., & Kalyan, A. (2023). *Anthropomorphization of AI: Opportunities and Risks*. CS ArXiv preprint.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*(12), 1845–1853. https://doi.org/10.1177/0956797610388814

Glikson, A. & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2). https://doi.org/10.5465/annals.2018.0057

Grice, H. P. (1957). Meaning. *The philosophical review, 66*(3), 377-388.

Guthrie, S. E. (1993). *Faces in the clouds: A new theory of religion*. Oxford University Press.

Hasan, A. (2023) *Why you are (probably) anthropomorphizing AI (Short Version)*. PhilArchive preprint.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science, 298*(5598), 1569-1579. https://doi.org/10.1126/science.298.5598.1569

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259. https://doi.org/10.2307/1416950

Hilton, D. (2007). Causal explanation: From social perception to knowledge-based causal attribution. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles*, 232–253. The Guilford Press.

Hockett, C. F. (1959). Animal "languages" and human language. *Human Biology, 31*(1), 32–39. http://www.jstor.org/stable/41449227

Hofstadter, D. R. (1995) *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* Basic Books.

Karataş, M. & Cutright, K. M. (2023). Thinking about God increases acceptance of artificial intelligence in decision-making. *Proceedings of the National Academy of Sciences.*

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual review of psychology, 31*(1), 457-501. https://doi.org/10.1146/annurev.ps.31.020180.002325

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature 521*, 436–444. https://doi.org/10.1038/nature14539

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.

McGlynn, J., & McGlone, M. S. (2019). Desire or Disease? Framing Obesity to Influence Attributions of Responsibility and Policy Support. *Health Communication, 34*(7), 689–701. https://doi.org/10.1080/10410236.2018.1431025

Oatley, K., and Yuill, N. (1985). Perception of personal and interpersonal action in a cartoon film. *British Journal of Social Psychology, 24*(2), 115–124. https://doi.org/10.1111/j.2044-8309.1985.tb00670.x

Piaget, J. (1926). *The Child's Conception of the World*. United Kingdom: Littlefield Adams Quality Paperbacks.

Schwitzgebel, E. & Shevlin, H. (2023). Opinion: Is it time to start considering personhood rights for AI chatbots? *Los Angeles Times*.

Song, Y., & Luximon, Y. (2020). Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors, 20*(18), 5087.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition.* Harvard University Press.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PloS one, 6*(2), e16782.

Thibodeau, P. H., Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PLoS One, 8*(1): e52961. https://doi.org/10.1371/journal.pone.0052961

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *The Washington Post.*

Troshani, I., Hill, S., R., Sherman, C. & Arthur, D. (2021) Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems, 61*(5), 481-491. https://doi.org/10.1080/08874417.2020.1788473

van Buren, B., Uddenberg, S., & Scholl, B. J. (2016). The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychonomic Bulletin and Review*, *23*(3), 797–802. https://doi.org/10.3758/s13423-015-0966-5

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Zhao, X., & Malle, B. F. (2022). Spontaneous perspective taking toward robots: The unique impact of humanlike appearance. *Cognition*, *224*. https://doi.org/10.1016/j.cognition.2022.105076

Intentional Action Without Action

Renee Rushing, Romy Vekony, and Michael Long

Florida State University

Introduction

Can one perform an intentional action without successfully performing it? On the face of it, this seems like a silly question. One can't intentionally do A without actually doing A. This is the intuition that many philosophers of action would share. For while the philosopher of action wants to determine what it is about an action that makes it intentional, the underlying assumption is that if we have an intentional action, we have an action. If, for example, we want to know what it is that makes Tallula's closing the door behind her an intentional action, the underlying assumption is that the door closed. Thus, many theorists (for instance, Knobe and Malle 1997, Mele 2001, Knobe 2003, Vekony et al. 2020) compose vignettes that feature a certain action, and then directly ask the participants to rate their agreement with a statement classifying the action as intentional. Due to the common assumption above, this method is seen as unproblematic.

However, if faced with evidence that ordinary English users are comfortable with saying that an agent 'intentionally' performed an action despite the agent's failure to successfully perform it, what lesson should philosophers of action engaged in experimental philosophy take? We designed a study to investigate whether there are instances in which participants would agree that an agent has performed an action, even though the agent's performance was unsuccessful. The results of our study indicate that English users are comfortable with saying that an agent intentionally performed an action despite the fact that the action did not occur. With the evidence garnered from our study, we posit two possible conclusions: either non-specialists are in vast error about the conception of intentional action, or the results garnered from experimental methods that ask directly for intentionality ascriptions are unreliable due to the flexible nature of ordinary language. We argue that the latter is the more reasonable conclusion, and suggest that a popular methodology in experimental philosophy of action should be re-evaluated.

Did A do X intentionally?

Many philosophers of action have an interest in understanding the folk conception of intentional action. Philosophers of action disagree about the features that are necessary for an action to be intentional, and they hope that discovering what the folk believe intentional actions consists of may help settle these disagreements. Since the concept of intentional action has its roots in everyday language and use, it would behoove philosophers not to stray too far from the source of the concept under their analysis, less they risk constructing or analyzing a different concept than they aimed for at the outset.

Sometimes, research into the folk concept of intentional action can result in startling evidence that a certain feature of the folk conception diverges greatly from the popular theories held by philosophers. In turn, philosophers have increasingly realized the value of determining the features of the folk conception of intentional action in its own right, and research has flowered as a result. Perhaps the most famous case where experimental methods have produced surprising results has been the cases run by Joshua Knobe (Knobe 2003, 2004, 2006). What is

now known as the "Knobe effect" is an apparent asymmetry in the folk concept of intentional action with regards to the moral valence of side-effects. People seem more ready to attribute intentionality to an unintended side-effect of an action if that side effect has negative moral valence (is morally wrong or harmful in some way), while they react in an opposite way to neutral or 'good' side effects.

Various explanations for this asymmetry have been proposed. One explanation is that folk intuitions on intentionality are directly tied to moral valence, or "how "bad" the outcomes seem." (Laurent et al 2020: 411) Knobe similarly argues that moral considerations figure into "the competencies people use to make sense of human beings and their actions" (Knobe 2010: 316). Another explanation is that folk hold multiple conceptions of intentional action. (Cushman and Mele, Cova et al 2012, Lanteri 2012, Mele and Cushman 2009.)

Much of the literature on folk conceptions of intentional action converges around cases featuring side-effects, but recent research (Vekony et al, 2020) has focused on folk concepts of intentional action in cases where there are no known bad side effects. Vekony et al set out to put the thesis that knowledge or awareness of one's action is necessary for acting intentionally to empirical test. Their research indicates that ordinary views require neither knowledge, nor awareness, of one performing an intentional action for one to have performed an intentional action.

However, many of these studies on folk conceptions of intentional action depend upon the assumption that philosophers can elicit folk assent or dissent about the intentionality of a particular cause by posing straightforward prompts using the term "intentionally." In this paper, we raise worries about the reliability of this method of interacting with participants. This method includes any experimental set-up using vignettes of an agent performing an action, with the participants being asked at least one question with the rough form of "Did A do X intentionally?" or being prompted to show their level of agreement with a sentence like "A did X intentionally."

This assumption about effective experimental methodology in turn rests on a deeper theoretical assumption about the nature of intentional action, one that runs through all of the philosophical literature on the subject. The assumption that for an agent to have performed intentionally, they must have performed an action is so basic that it seems trivial and not worth mentioning. If one holds this assumption, as the majority of theorists do, it is natural to assume that the 'folk' hold it as well. The idea that a direct question/prompt asking about the 'intentionality' of an action using the term "intentionally" can elicit reliable evidence about the folk conception of intentional action follows naturally: if the folk recognize an action being performed, we can ask them if they categorize it as intentional or not. However, as we shall show, the situation is complicated if people are willing to attribute 'intentionality' to an action that actually hasn't been performed.

Here are a few examples of experimental work on intentional action that feature this method. The first example is pulled from Knobe (2003) "Intentional Action and Side Effects in Ordinary Language.":

> Each subject was randomly assigned to either the 'harm condition' or the 'help condition'. Subjects in the harm condition read the following vignette:
>
> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the

> environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed.
>
> These subjects were then asked to determine how much blame the chairman deserved for what he did (on a scale from 0 to 6) and to say whether they thought the chairman intentionally harmed the environment. (Knobe 2003: 190)

In this study, participants were asked a categorical yes/no question with regards to the intentionality of CEO's action. We see something similar in Lanteri (2009). After presenting his re-worked version of Knobe's CEO cases, Lanteri asked participants the following two questions: "According to you, the CEO harmed/helped the environment… (a) intentionally (b) not intentionally (c) neither According to you, for his decision, the CEO should be… (a) praised (b) blamed (c) neither." (Lanteri 2009: 718)

A different sort of example can be found in Mele & Cushman (2007)'s report. After presenting their vignettes, instead asking a categorical question, they asked participants to rate their responses on a 7-point Likert scale (1 being a strong no, 7 being a strong yes) to prompts such as "Did Lydia intentionally hit the bull's-eye?" and "Did Tom intentionally hit the red striped ball into the side pocket?" (Mele & Cushman 2007: 186)

Vekony et al 2020 display a similar method of probing for intentionality ascriptions. The authors ran a series of vignettes like this:

> [Basketball] Andy is a 92% free throw shooter. One evening, he is at the gym practicing his free throws. He lines up and takes the shot. But just as the ball leaves his hands, lightning strikes the building. The power goes out and it is pitch black. There is also a loud clap of thunder. Due to this, Andy could not see or even hear whether he made the shot. He is completely unaware of whether he sank the shot. But he did in fact sink the shot." (Vekony et al 2020: 5.)

Vekony et al then presented three randomized test statements and asked participants to rate the level of their agreement on a 7-point Likert scale, 1 for "strongly disagree" and 7 for "strongly agree".

> When Andy was sinking the shot, he knew that he was sinking it. (Knowledge)
>
> While Andy was sinking the shot, he was aware of sinking the shot. (Awareness)
>
> Andy intentionally sank the shot. (Intentionality) (Vekony et al 2020: 5)

Multiple other studies like this can be found, where participants are asked directly to either rate their agreement with a sentence that directly ascribes intentionality to an agent's action, or to answer a question asking directly about whether an agent performed an action intentionally. It is clear that a good portion of experimental philosophy of action relies on this kind of method of probing folk intuitions. And indeed, it seems like the most straightforward way to do it. However, despite its simplicity, it may not be as effective as repeated utilization by researchers suggest it might be. One familiar with the fact that in ordinary language pragmatic context often alters the meaning of a term may worry that participants are cued by seemingly irrelevant features of vignettes into altering the context in their minds, or even supplying their own background context when presented with the sparse context of an experimental survey.

We decided to test the reliability of this method. Our contention is that if we find that even performing the relevant action isn't necessary for people to agree with statements to the effect that "X performed A intentionally," then we cannot expect *any* experiment that asks participants in such a manner to produce results that can reliably inform us of the ordinary concept of intentional action.

Our study and methods

For our study, we devised a total of ten vignettes. Five of them featured actions that were successfully performed, and the other featured unsuccessful attempts to perform the action. 390 participants were recruited from Amazon Mechanical Turk (https://www.mturk.com) and tested in Qualtrics (https://www.qualtrics.com) We randomly sorted the participants into control and test groups and provided them with five vignettes each. The cases of successful action served as our control cases. The cases of unsuccessful action were our test cases. We wanted to know whether there would be a significant difference between the intentionality ascriptions between the control and test cases. Our hypothesis was that there would be no significant difference; people would be willing to ascribe intentionality to the unsuccessful cases of action at similar rates as for the successful cases.

Below is an example of a vignette we ran, both its successful completion and unsuccessful forms:

> [Car Lock Success] Tallula's car has a glitch that she is unaware of. When she arrives at work, Tallula clicks the automatic lock button on her set of keys to lock her car doors. Her car doesn't make the usual beep. So Tallula thinks the car didn't lock; and because she is in a hurry, she walks away. But, in fact, the car doors locked.
>
> [Car Lock Fail] Tallula's car has a glitch that she is unaware of. When she arrives at work, Tallula clicks the automatic lock button on her set of keys to lock her car doors. Her car makes the usual beep, but – although Tallula doesn't know it – the car remains unlocked.

After being presented with a vignette, participants were then presented with these instructions in a randomized order:

> Using a scale from Strongly Disagree (1) to Strongly Agree (7), indicate your agreement with each item.
>
> Tallula intentionally locked her car doors.
>
> Tallula tried to lock her car doors.
>
> Tallula intended to lock her car doors.

We also included an attention check in the form of a yes/no question:

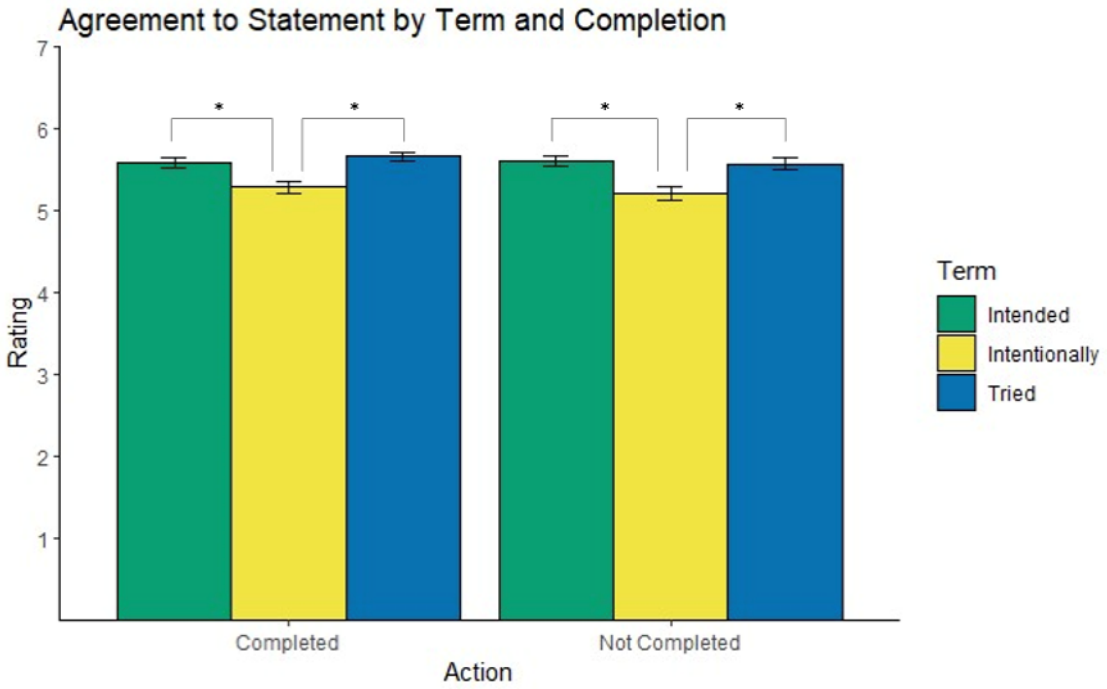
> Did Tallula's doors remain unlocked? (yes or no)

The attention check was crucial for our study. It was important that they register that the relevant action was successfully completed/unsuccessful (i.e. that the door was locked/unlocked, and so on for the rest of the vignettes). We decided the most neutral way to ask participants was to ask about the physical status of the object the agent acted upon in each vignette, rather than ask a question like "Did Tallula successfully complete the action?" because we worried that this might be interpreted as another question about the intentionality of the relevant action. The responses of the participants that failed the attention check were not included in the final data.

In addition, it was important that we included response options like *intended to* and *tried to* so that people had adequate ways of expressing themselves. We didn't want them to respond in agreement with the *intentionally* condition simply as a way of trying to express that they thought Tallula only intended to or tried to lock her door.

Results

Below is a graph of our results in Figure 1.



Agreement to statement about action on a scale from 1 to 7 (1-Strongly Disagree, 4-Neither Agree nor Disagree, 7-Strongly Agree). Asterisks represent significant differences in ratings. Error bars represent standard error. n = 160.

Overall, there was a positive response (neutral response = 4) that all terms accurately described all actions across vignettes and completion of the action. Participants agreed that the agents in the vignettes *intentionally*, *intended to*-, and *tried to* complete the action, whether the actions were completed or not. *Intentionally* was consistently rated lower than the other two terms across conditions ($ps < .05$), suggesting that participants understood the term to have a different meaning than the other two terms. *Intended* and *tried* were not rated different from one another. This was the same across the vignettes, where *intentionally* was rated lower ($ps < .05$), or marginally lower ($ps < .1$) than the other two terms which were not different from each other ($ps > .05$).

There were no differences in the ratings between the group of participants who read about a scenario in which the action was completed and the group that read about actions that were not completed ($ps > .05$).

While we do see a lower rating for *intentionally* across both groups, it is important to note that there were no differences between groups on this variable. In addition, both groups were

willing to rate *intentionally* higher than a neutral response, even when offered the options of *intended to* and *tried to*. So, the difference between whether an action was successfully completed does not suggest itself as a reason for why participants were reporting intentionality lower.

Conclusion

So, what are we to conclude? One could argue that the folk are just completely mistaken. This is always an option one can take if one sees experimental results that put folk responses in contradiction with one's own theory. But here, the folk seem to be in direct contradiction to an underlying assumption about intentional action that seems so basic that the thing that the error theorist is forced to conclude is that ordinary English speakers don't know what an action is at all. They simply cannot be relied upon to recognize when an action has been performed and when it hasn't.

This sort of accusation of folk error goes perhaps deeper than we may want to go. The other conclusion we can come to is that ordinary English speakers do not use the word 'intentionally' in a manner consistent with how many philosophers of action assume they would. And that this doesn't necessarily mean any party is in error on their conception of intentional action, but that the particular method of asking participants whether they think an action is 'intentional' or not is not the best way of actually getting at the folk conception of intentional action.

Further Discussion

We're not going to speculate much in this paper about why we are seeing the results that we are seeing. We will leave that to a further project. Our project was aimed at testing whether a certain method of probing for folk responses shows a disagreement with a fundamental assumption shared by most philosophers of action.

However, it is worth noting an asymmetry within our vignettes. In each 'successfully completed' scenario, we describe the agent as believing that their action was unsuccessful, while in the 'uncompleted' scenarios we describe them as being unaware of whether their action was completed. A large concern of ours when writing the vignettes was to provide a plausible story for our participants. Not only did we want them to be every-day, relatable actions so as to better illicit gut reactions, we wanted to make sure our agents were not acting in unintuitive ways, such as, for instance, believing that the car is not locked when receiving the 'beep' (normally taken as evidence it is locked), or not having a reason (such as being in a hurry) to fail to double check the locks upon not receiving a 'beep.' We didn't want participants to fill in background context themselves in an effort to make sense of an unintuitive story.

Looking back, we realize that the representations of the agents' mental states in each scenario may be a confounding factor in our study. Perhaps, one could say, the difference in the agents' beliefs about whether they have successfully performed an action or not have an effect on the people's intuitions great enough to effectively neutralize any effect that could be created by the difference in the completion conditions alone. This may be worth looking into with a follow-up study. But for now, we believe our results still support our contention that direct prompts featuring the word "intentionally" do not prove to be a reliable method for getting at the folk concept of intention.

Bibliography

Adams, Fred & Steadman, Annie (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? Analysis 64 (2):173–181.

Cova, F., Dupoux, E., & Jacob, P. (2012). On doing things intentionally. Mind & Language, 27(4), 378-409. https://doi.org/10. 1111/j.1468-0017.2012.01449.x

Cova, Florian (2013). Unconsidered Intentional Actions. An Assessment of Scaife and Webber's 'Consideration Hypothesis'. Journal of Moral Philosophy (1):1-22.

Cushman, Fiery & Mele, Alfred (2008). Intentional action : two-and-a-half folk concepts? In Joshua Michael Knobe & Shaun Nichols (eds.), Experimental Philosophy. Oxford University Press. pp. 171.

Hindriks, Frank (2014). Normativity in Action: How to Explain the Knobe Effect and its Relatives. *Mind and Language* 29 (1):51-72.

Knobe, J. (2003). Intentional action and side effects in ordinary language. Analysis 63 (3):190-194.

Knobe, Joshua (2003). Intentional action in folk psychology: An experimental investigation. Philosophical Psychology 16 (2):309-325.

Knobe, J. (2010). Person as scientist, person as moralist. Behavioral and Brain Sciences, 33(4), 315-329. https://doi.org/10.1017/ S0140525X10000907

Lanteri, A. (2012). Three-and-a-half folk concepts of intentional action. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 158, 17-30. https://doi.org/10. 2307/41406995

Lanteri, A. 2009. Judgements of intentionality and moral worth: Experimental challenges to Hindriks The Philosophical Quarterly 59(237): 713–720.

Laurent SM, Reich BJ, Skorinko JLM. Understanding Side-Effect Intentionality Asymmetries: Meaning, Morality, or Attitudes and Defaults? Pers Soc Psychol Bull. 2021 Mar;47(3):410-425. doi: 10.1177/0146167220928237. Epub 2020 Jun 29. PMID: 32597329.

Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. Midwest Studies in Philosophy, 31(1), 184-201. https://doi.org/10.1111/j.1475-4975. 2007.00147.x

McGuire, John Michael (2012). Side-effect actions, acting for a reason, and acting intentionally. Philosophical Explorations 15 (3):317 - 333.