Filozofia Nauki (*The Philosophy of Science*)
ISSN 1230-6894 e-ISSN 2657-5868
2021, vol. 29(2) [114]: 209-227
DOI: 10.14394/filnau.2021.0011

JOANNA KOMOROWSKA-MACH,* ANDRZEJ SZCZEPURA**

FIRST-PERSON AUTHORITY THROUGH THE LENS OF EXPERIMENTAL PHILOSOPHY***

Abstract

In this paper, we analyze the problem of first-person authority and the possibility of disagreement over mental states between first- and third-person ascribers. We explain why discussion on this matter should be preceded by empirical study on the actual strength, scope, and restrictions to such authority. We present a new study in which we show that the type of the ascribed mental state and the kind of interpersonal relationship between speakers both influence the strength of first-person authority. We also suggest that analysis of a disagreement between a first- and a third-person ascriber of a mental state should take into account the intuition that it is possible that neither of these disagreeing speakers is wrong in their ascriptions.

Keywords: first-person authority, self-ascriptions, mental states ascription, privileged access, disagreement

INTRODUCTION

The problem of self-knowledge, like many other problems of the philosophy of mind, has undergone a rapid naturalistic shift in recent decades. First-person authority — a special status that our self-ascriptions of mental states enjoy in socio-linguistic practice, and that for centuries was understood as a

* Faculty of Philosophy, University of Warsaw, Krakowskie Przedmieście 3, 00-927 Warsaw, Poland, e-mail: j.komorowska-mach@uw.edu.pl, ORCID: https://orcid.org/0000-0002-8287-6668.

** Faculty of Philosophy, University of Warsaw, Krakowskie Przedmieście 3, 00-927 Warsaw, Poland, e-mail: a.szczepura@student.uw.edu.pl.

*** We want to thank Bartosz Maćkiewicz for his help with implementation of the studies and statistical analyses, and Katarzyna Kuś for the support of the project and insightful remarks on the paper.

mere symptom of epistemically privileged access to one's own mental state — has come to be treated as a phenomenon that itself is worth investigating. Philosophers such as Crispin Wright (1998), Richard Moran (2001), Dorit Bar-On (2004), David Finkelstein (2003), and Peter Carruthers (2011) (to name just a few) reject the claim that we can "see" our mental states in a specially secure way, and at the same time they try to do justice to the intuition that our self-ascriptions of mental states have a special status among other kinds of utterances. The problem that these anti-introspectivist philosophers try to solve is, therefore, how to explain first-person authority without postulating special epistemic access to our mental states. However, there is a more general problem with this approach: the phenomenon of first-person authority itself is very vaguely described, and even among naturalistically oriented philosophers there is no agreement on its most fundamental features.

In this paper, we show how the empirical perspective can be useful in tackling this problem. The proposed approach deviates from the most common usages of the paradigm of experimental philosophy in the philosophy of mind. Typically, there are two ways in which experimental philosophy is engaged in solving the puzzles of the philosophy of mind. The first (which roughly falls under the category of a "positive program of experimental philosophy," as distinguished by Stephen Stich and Kevin Tobia (2016)) is the empirical research of philosophical intuitions that may inform the conceptual analysis of crucial notions such as consciousness, mind, belief, etc. The second (as an instance of a "negative program") examines trustworthiness of intuitions that philosophers rely on by testing their sensitivity to factors that should not be relevant to the truth and falsity of philosophical claims (ethnicity, gender, affectivity, presentation order, etc. (Weinberg, Alexander 2014)). In both cases, empirical findings are supposed to serve as validation of the effects of philosophical work. Our approach is different. We use the tools of experimental philosophy not to explain first-person authority but to describe it more precisely and dispel some controversies before the explanatory part of the study can proceed. There is no clear consensus on the strength and scope of first-person authority. Also, little has been said about exceptions to it — that is, situations in which the self-ascription of a subject is overtly questioned by her interlocutor. In our opinion, relying solely on armchair intuitions and subjective observations in this matter may lead to unreliable and inconsistent results.

To formulate the problem of the strength and scope of first-person authority in a way that is suitable for empirical study, we focus on a special kind of disagreement between two speakers: one who ascribes a state to herself and a second who denies this ascription from the third-person perspec-

tive. In section 1, we describe this kind of disagreement as an exception to the first-person authority of self-knowledge. In section 2, we present an empirical study in which we evaluate some hypotheses concerning first-person authority. Finally, we discuss our results and explain the impact they may have on theoretical studies on the matter. We also propose some directions for further empirical research.

1. FIRST-PERSON AUTHORITY AND ROOM FOR DISAGREEMENT

Let us imagine John and Sue talking. John says, "I am sad," and Sue replies, "No, you are not sad." At first glance, it is clear that only one of them can be right: John is either sad or not. If forced to make a choice, we would probably trust John rather than Sue in this matter. However, most of us would also agree that it is not impossible for John to be wrong: for some reason, he may fail to grasp his own state adequately, and Sue's reaction may even help him to correct his judgment in such a case.

In this paper, we use the term "first-person authority" in a broad and neutral way to refer to the special status of self-ascriptions: utterances in which the subject uses categories of folk psychology to ascribe to herself an occurring mental process or a state. "First-person authority" refers to the fact that self-ascriptions are rarely questioned or corrected, and subjects who avow their own mental states are (unlike third-person observers) not asked for additional justification for their statements. For example, if somebody utters "I am excited," "I feel pain," or "I hope he will be here soon," we do not normally ask, "How do you know this?" or "Are you sure?"; rather, we take such utterances for granted. First-person authority may but does not have to stem from the epistemic privilege of self-knowledge (see also Davidson 1984, Bar-On 2004). The broad characterization of first-person authority that is proposed here is noncontroversial for both introspectivists and proponents of alternative approaches to self-knowledge. As Derek Jongepier and Fleur Strijbos state it,

> It is important to realize that epistemic privilege and first-person authority are two different explananda, each allowing, in principle, for different explanations. Roughly, the first question concerns the epistemology of our self-reports — what makes these items particularly knowledgeable? — whereas the second concerns the question of what underlies our practice of taking each other at our words. (Jongepier, Strijbos 2015: 124)

The idea that is followed by anti-introspectivists — namely, shifting attention from epistemic privilege to first-person authority — may be considered similar

to the more recent yet better-known proposition of David Chalmers (2018) that we should rethink the hard problem of consciousness in the context of the meta-problem of consciousness by asking why we have the intuition that consciousness is problematic in the first place. In our case, the "meta-problem of self-knowledge" could be formulated like this: why do we have the intuition that ascribing occurring mental states to ourselves has a special status that needs to be explained.[1] Our approach in this paper takes one more step back: what precisely are these unexplained intuitions that we have about our self-ascriptions?

Early versions of introspectivism saw self-ascriptions as always correct (due to the assumption that the faculty of introspection is direct and infallible). Contemporary philosophy of mind generally admits that first-person authority has its exceptions: we do not always treat self-ascriptions as correct and infallible. However, the nature and scope of these exceptions remain undefined and tend to be the object of controversy between philosophers. If first-person authority is the tendency to take first-person ascriptions for granted, exceptions to it should be seen as situations in which a first-person ascription is doubted or challenged. The observable symptoms of such an attitude are cases of disagreement between first- and third-person ascribers of a mental state — that is, cases in which the interlocutor overtly denies the self-ascription of the first speaker.[2]

Psychotherapeutic discourse (or at least its pop-culture version) is commonly used as an exemplary context of situations in which a first-person ascription of a mental state is openly challenged by the interlocutor ("I am not angry at my father!," "Are you sure about it?"). In everyday conversations, first-person authority does not seem absolute either: it is not surprising or improper if someone disagrees with her interlocutor's self-ascription. If John says, "I am sad," Sue's response "You're not sad, you're just tired" is neither absurd nor linguistically incorrect (for more examples, see Bar-On 2004: 99, Finkelstein 2003: 192-193, Schwitzgebel 2008: 252). However, it remains unclear in which cases we find disagreement between a first- and third-person ascriber natural and acceptable. We have singled out three specific problems concerning this matter that we find suitable for an empirical study.

---

[1] For an alternative view on the relation between these two problems, see Schwengerer 2019.

[2] Note that there are two general types of situations in which we can deny someone's self-ascriptions: cases in which we believe that our interlocutor has made a mistake in her self-ascription, and cases in which we suspect her of lying. We only see the first situation as a genuine exception to first-person authority, and we focus only on sincere self-ascriptions (more about this distinction in Rodriguez 2012).

A. Is first-person authority homogenous for different types of ascribed mental states?

Wright's armchair observation is that first-person authority is stronger for phenomenal than for intentional states. He contrasts utterances in which we self-ascribe phenomenal states, such as sensations and emotions, with those in which we self-ascribe intentional states such as beliefs and intentions; he argues that the latter are more often questioned or corrected (Wright 1998: 17). This intuition is supported by philosophers who propose pluralistic models of self-knowledge and provide separate explanations for the first-person authority of intentional and phenomenal states (see, e.g., Boyle 2009, Coliva 2016) or by those who restrict their models to just one of these types (e.g., Moran 2001, Gertler 2001, Nichols, Stich 2003, Goldman 2006). Contrary to Wright's claim, Bar-On argues that both phenomenal and intentional mental states enjoy a special status of the same type (Bar-On 2004: 5-6), and many other approaches to self-knowledge (both introspectivist and non-introspectivist) disregard the difference between them, even if they do not deny it explicitly (e.g., Russell 1912, Finkelstein 2003).

We believe that this discussion, which underlies further theoretical work on first-person authority, may be informed by experimental research. If Wright is correct, a disagreement between first- and third-person ascriptions concerning phenomenal states should be considered less adequate or be more often resolved in favor of the first-person ascriber than a disagreement over self-ascriptions concerning intentional states.

B. Which pragmatic features of a disagreement between a first- and a third-person ascriber influence the strength of first-person authority?

The default view on first-person authority is that it is a symptom of the epistemically privileged access that every person has to their own mental states; however, if first-person authority is only grounded in the epistemic privilege of self-knowledge, others' questioning of it should be seen as inadequate or rare in all cases, independently of the context.

As philosophers such as Finkelstein (2003) and Lukas Schwengerer (2019) notice, we find the questioning of self-ascriptions to be more common and appropriate in some social situations than in others. Apart from the context of psychotherapy, these two philosophers claim that the typical cases of disagreement are those in which disagreeing interlocutors know each other well and enjoy a close relationship. The fact that the type of the relationship between speakers influences the strength of a first-person disagreement may therefore be seen as an argument against a purely epistemic interpretation of first-person authority.

C. What is the status of a disagreement about mental states between first- and third-person ascribers?

It is not obvious whether we should treat conflicting first- and third-person ascriptions (e.g., "I'm sad," "No, you're not sad, you're just tired") as two contradicting propositions (as in "It's a cat," "No, it isn't a cat. It's a dog") or as analogous to predicates such as "tall" or "pretty," whose meaning is vague or relative to the speaker. The question is, therefore, whether a disagreement between a first- and a third-person ascriber of a mental state is seen as necessarily resolvable in favor of one of the speakers (one of them is right and the other is wrong) or whether there is a possibility that neither of the speakers is wrong, despite the contradictory relation between their statements.

These three questions can hardly be answered from the armchair and — to our best knowledge — there has been no previous empirical research on these matters. Therefore, we decided to use the methods of experimental philosophy to investigate laypeople's intuitions on these three matters.

2. EXPERIMENTAL STUDY

In this study, our goal was to explore the determinants of the strength of first-person authority and the tendency to see an argument between first- and third-person ascribers as a situation in which it is possible for neither of the speakers to be wrong, despite the contradictory character of their statements. We decided to evaluate two hypotheses:

H1. First-person authority for phenomenal self-ascriptions is stronger than for intentional self-ascriptions.

H2. In close interpersonal relationships between interlocutors, first-person authority is weaker than in relationships that are not close.

Additionally, we wanted to take into account the question concerning the status of disagreements between first- and third-person ascribers. Since theoretical discussion on this matter is very limited and, in our opinion, does not allow us to formulate straightforward hypotheses, we decided to treat this part of the research as exploratory.

2.1. MATERIALS AND METHOD

Participants. 661 respondents took part in the study, all of whom were recruited via the Amazon MTurk platform. Participants were paid $0.50 each

for participation in our survey. 60 of them were excluded from the study as they did not pass the attention test, therefore the answers from only 601 participants were analyzed. The study was conducted in the form of a survey on the online LimeSurvey software.

Experimental model. The study was performed in a 2*2*2*2 between-subjects design. Each subject was randomly assigned to one of sixteen groups. For each group, one possible combination of the values of four independent dichotomic variables was assigned.

The main independent variables were:

IV1. The type of internal state ascribed by speakers (a belief as an example of an intentional state, and an emotion as an example of a phenomenal one).

IV2. The type of relationship between speakers (a close friend or a new colleague).

The extraneous independent variables were:

IV3. Positive or negative character of the internal state.

IV4. The order in which the examined person was presented with possible options (with the first-person ascriber located on the extreme left or extreme right of the scale).

In all groups, two dependent variables were measured. The main dependent variable was DV1: the subject's intuition about the strength of the first-person ascription in the presented case, taking values from "the first-person ascriber is definitely right" to "the third-person ascriber is definitely right." The additional dependent variable was DV2: the subject's intuition about the status of the disagreement in the presented case, taking one of two values (speakers may or may not be both wrong in the presented case of disagreement).

The subjects were also asked several meta-questions about their feelings about the story and about the questions presented in the main part of the study.

Materials and procedure. Respondents were first acquainted with a short story whose content was determined by the group to which the respondent belonged (see Appendix A for all the variants of the story). In the story, two people (Tom and Ben) argued about Tom's internal state. Question 1, which served as indicator of variable IV1, was intended to determine which of the speakers was right according to the participants (see Appendix B). The

respondents were asked to mark their answer on a 7-point Likert scale,[3] where 1 stood for "Definitely Ben" and 7 stood for "Definitely Tom" — or conversely (depending on the value of V4).

For example, in group No. 13, which got the version of the task with the "intentional" value of IV1, the "colleagues" value of IV2, the "negative" value of IV3 and the "Ben first" value of IV4, the story and the question looked like this:

> *Suppose Tom is talking to his new colleague, Ben. Tom sincerely says, "I believe it's going to rain tomorrow." Ben disagrees, "No, you do not." Which of them is right?*
>
> *1 2 3 4 5 6 7*
> *Definitely Ben — Hard to say — Definitely Tom*

We explicitly stated that Tom is speaking sincerely, in order to rule out the interpretation that Ben disagrees with him because he believes that Tom is lying.

If a respondent marked the middle point ("Hard to say") on the Likert scale, she received a supplementary question asking why she chose this option (with three possible answers: "I think they are both equally right," "I don't know whether Tom or Ben is right" or "I found the question vague"; Appendix C). This was done to separate answers that indicated that the subject had a problem understanding or interpreting the story or the question. Both "I don't know whether Tom or Ben is right" and "I found the question vague" answers were excluded from later analysis (respectively 77 and 26 answers of these types).

After answering the first question, the respondents from all groups were presented with Question 2: "Is it possible that, despite their disagreement, neither of them is wrong about Tom's beliefs/feelings?" Question 2 was intended to serve as an indicator of variable IV2 (Appendix D). Possible answers to this question were "yes" and "no". The structure of the survey forced respondents to answer this question.

Subsequently, the respondents were asked about their general opinion of the story, and both previously answered questions (see Appendix E). For each of the five questions, the possible answers were "yes" or "no." The reason for these questions was to determine whether respondents found the story, or either of the two questions, inadequate or hard to understand.

After answering the above questions, the subjects passed a short attention test (Appendix F). The goal of the test was to screen out participants who had

[3] In the pilot study, we tested the continuous scale, but we found a significant effect of the order of the outermost answers. The problem did not occur in the version of the study with the Likert scale presented here.

not read the instructions. As stated in the section "Participants" above, this excluded 60 respondents from the analysis.

The last questions asked whether they had completed at least five courses in philosophy as part of their education and whether their native language is English (both with "yes"/"no" answers).

As a result of this process, we obtained a complete set of answers from each respondent:

a) an answer to the first question, "Which of them is right?";

b) an answer to the second question, "Is it possible that, despite their disagreement, neither of them is wrong about Tom's beliefs/feelings?";

c) five "yes"/"no" answers to the meta-questions about the story and two main questions asked before;

d) answers to the questions about philosophical education and native language.

Additionally, from the respondents who chose the "Hard to say" answer to the first question, we received information about the reason they selected this particular answer.

As for the first answer, because the order of the answers on the Likert scale (V4) was counterbalanced, we had to recode half of the raw answers. As a result, for all speakers the answer "1" means that it is the first-person ascriber who is definitely right, and the answer "7" means it is the third-person ascriber who is definitely right.

2.2. RESULTS

Aside from the experimental questions, we controlled participants' philosophical education and whether or not English was their native language. We found a negligible correlation between philosophical education and the answers to the main questions ($r = 0.147$; $p < 0.001$; $n = 661$), and there was no significant correlation between language status and these answers ($r < 0.001$; $p = 0.982$; $n = 661$). Therefore, we decided not to exclude non-native English speakers or subjects with philosophical education from the study. Controlled variable IV4 (the order in which the possible answers were presented to the participant on the Likert scale) did not influence the answers in a significant way ($r = -0.023$; $p = 0.578$; $n = 601$).

The presence of first-person authority was visible in the overall results of the survey: the distribution of the answers to the first question differed sig-

nificantly from the uniform distribution (M = 2.481; U= 102383; p < 0.001; n = 601).

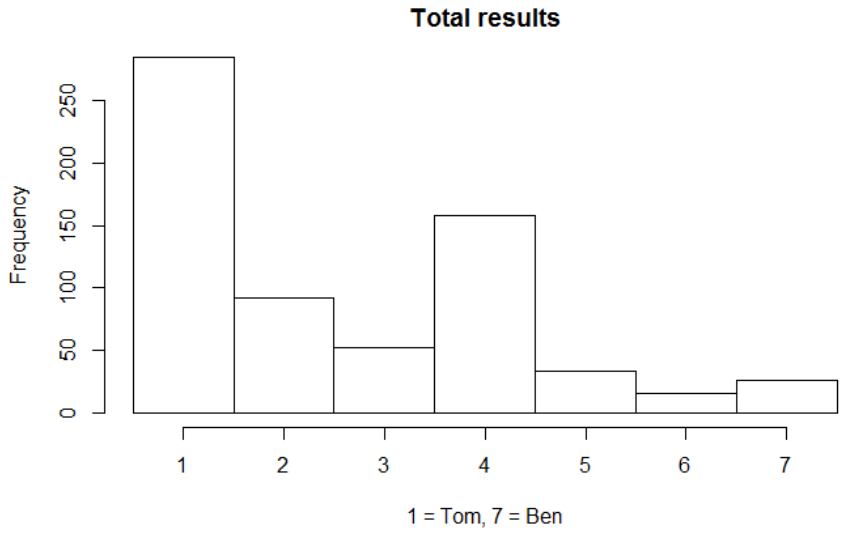


Fig.1. Results concerning a general first-person authority effect: answers to the question "Which of them is right?," with "hard to say" answers included

The presence of first-person authority was also confirmed by comparing the answers to two additional questions: only 9.5% (57/601) of participants answered "yes" to the question concerning whether what the first-person ascriber says in the story "is strange," compared to 66.9% who answered "yes" (402/601) in the analogous question about the utterance of the third-person ascriber ($\chi^2(1) = 417$; $p < 0.001$). At the same time, subjects did not see the authority of first-person ascribers as absolute: only less than half of them (45.1%) chose answer 1 (Tom is definitely right).

Hypothesis H1. As for the first of our hypotheses, H1, we found a significant effect with regard to the difference in the strength of first-person authority in scenarios concerning phenomenal and intentional self-ascriptions ($M_{ph} = 1.873$; $M_{int} = 2.556$; U = 23142; $p < 0.001$; n = 601), with first-person authority being stronger in phenomenal cases. What stands out is the significant level of "Hard to say" answers in scenarios with ascriptions of intentional states: 115 compared to 25 in cases of ascriptions of phenomenal states.

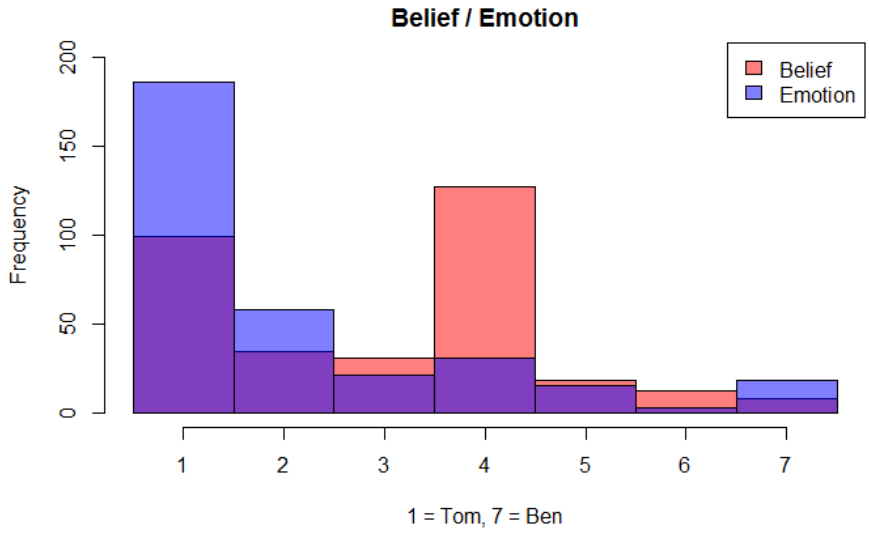


Fig. 2. Results grouped by the values of variable IV1: the type of a mental state ascribed

The effect is present both when considering only scenarios in which the interlocutors are close friends (M_{ph} = 2.028; M_{int} = 2.589; U = 6066; p = 0.002; n = 299) and when considering only scenarios in which they are merely colleagues (M_{ph} = 1.718; M_{int} = 2.523; U = 5518; p < 0.001; n = 302).

Hypothesis H2. As for the second hypothesis, H2, we found that in closer interpersonal relationships between interlocutors, the third-person ascriber is more likely to be seen as being right than in less close relationships (M_{friend} = 2.269; $M_{colleague}$ = 2.065; U = 33622; p = 0.037; n = 601).

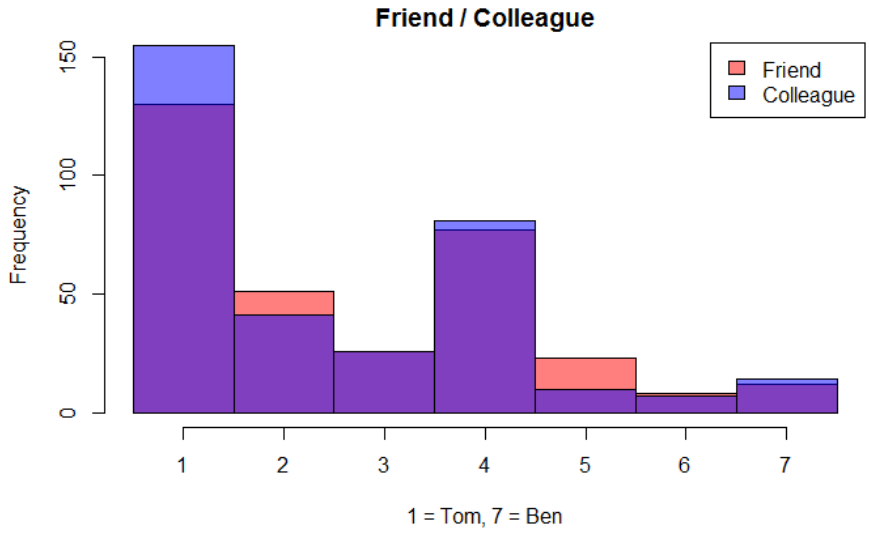


Fig. 3. Results grouped by the values of variable IV2: the type of relationship between speakers

However, when testing "feel" and "belief" types of scenarios independently, this effect was statistically significant only for the scenarios with phenomenal self-ascriptions (M_{friend} = 2.028; $M_{colleague}$ = 1.718; U = 11303; p = 0.021; n = 301). It was not present in scenarios concerning intentional self-ascriptions (M_{friend} = 2.589; $M_{colleague}$ = 2.523; U = 5916; p = 0.329; n = 300).

The status of disagreement. We found that almost half of the subjects (260 out of 601, 46.6%) were willing to say that in cases of disagreement between first- and third-person ascribers, it is possible that neither of the speakers is wrong. This tendency was higher in scenarios concerning propositional self-ascriptions than in scenarios with phenomenal self-ascriptions (49.3% and 43.9% of participants in each group, respectively), but the difference turned out to be statistically insignificant ($\chi^2(1) = 1.6$; p = 0.206; n = 601).

Other controlled variables. Variable V3 (the positive/negative character of the internal state) did not influence the answers in any significant way (M_{pos} = 2.162; M_{neg} = 2.171; U = 31392; p = 0.789; n = 601). It had no significant effect when taking into consideration only the answers gathered from the propositional scenarios (M_{pos} = 2.431; M_{neg} = 2.67; U = 5225; p = 0.26; n = 300), or only the ones from the phenomenal scenarios (M_{pos} = 1.972; M_{neg} = 1.77; U = 11186; p = 0.064; n = 301). However, a trend can be seen in the latter: par-

ticipants were more willing to agree with the first-person ascriber in scenarios in which a self-ascription concerned a negative phenomenal state.

As described previously, in all scenarios the respondents had the option to give one of three sub-types of a "Hard to say" answer: "I think they are equally right," "I don't know whether Tom or Ben is right," or "I found the question vague." To determine whether any of our independent variables influenced the understandability of the question or the tendency to avoid the answer, we checked the way in which the respondents who chose to answer "Hard to say" answered the sub-question, depending on the values of the variables IV1, IV2, and IV3 in their version of the task. We ran a series of Fisher's exact tests and found no significant difference in the distribution of the answers to the sub-question between the groups and all the gathered data.

2.3. DISCUSSION

Our participants generally did not see first-person authority as absolute. Less than half had the intuition that a first-person speaker is definitely right, despite having no information about the third-person ascriber's reason to disagree. This result is consistent with the intuition shared by most contemporary philosophers: it is neither absurd nor improper to question another person's self-ascription. It may, however, be a challenge to those who claim that self-ascriptions cannot be reasonably questioned. This may be true not only about classic introspectivism (it would be difficult to identify a contemporary supporter of this view) but also about approaches that claim that the special status of self-ascriptions arises from the fact that they constitute or contain ascribed states (see, e.g., Moran 2001, Burge 1988).

The results of our study confirm our first hypothesis — namely, first-person authority for phenomenal self-ascriptions is stronger than for intentional self-ascriptions. This result should be addressed by philosophers who propose a unified explanation of first-person authority, as they seem to assume that the explained phenomenon may be described as homogenous for different kinds of ascribed mental states.

The second hypothesis was confirmed partially: first-person authority is significantly weaker in close interpersonal relationships than between relative strangers, but only when we analyze the ascription of phenomenal states. This effect may be interpreted as suggesting that the difference between first-person authority concerning phenomenal and intentional states might be not only quantitative (as shown by the confirmation of H1) but also qualitative. Such an interpretation is also in line with approaches to first-person authority that seek different explanations for our self-knowledge concerning phenomenal

and intentional states, such as the pluralist accounts offered by Annalisa Coliva (2016) or Matthew Boyle (2009).

However, it is important to notice that in our materials we used only two examples of phenomenal states and two examples of intentional states. To be able to draw general conclusions, more types of mental states from these two categories should be included.

We showed that a surprisingly high percentage of participants (almost half of them) were willing to admit that it was possible that neither of the speakers is wrong despite the explicit contradiction between the sentences they uttered. This result may be seen as an argument against interpreting disagreement between first- and third-person ascriber as analogous to genuine disagreements concerning objective facts. This may be interpreted in a relativist manner by claiming that disagreements between first- and third-person ascribers are only superficial, as in fact both speakers use mental predicates in different ways (see MacFarlane 2014, Kölbel 2004). However, this effect may also have a simpler explanation if we interpret it merely as information that mental predicates are vague and that a discussion about whether somebody is sad or not may be analogous to a discussion over whether somebody is tall or not. Another question is whether participants interpreted the question "Is it possible that neither of them is wrong?" as stating that none of the speakers has a false belief and not as stating that none of them did something wrong in a moral sense. Furthermore, there are two possible problems with the way we formulated the question in the "belief" scenario.[4] First, the belief in question ("It is going to rain tomorrow") concerns contingent future events, which could make its truth-value vague for some of our participants. Second, when asked whether Tom was right in saying "I believe that it is going to rain tomorrow," some of the participants might have understood it as a question about him being right not about his mental state but about tomorrow's weather.

To clarify these issues, further investigation of the matter is needed. In order to rule out the possibility that this effect is not characteristic of disagreement between first- and third-person ascribers of mental states but applies more generally to all discussions about mental states, this seemingly large effect should be compared with cases of disagreement between two people who ascribe a mental state to somebody else from the third-person point of view. In order to resolve doubts as to the validity of the "belief" version of the study, we should use scenarios concerning beliefs about present or past events and ask a more precise question about them ("Which one of them is right about Tom's belief?"). We hope to address these issues in future studies.

---

[4] We are grateful to the editors for pointing out these problems.

CONCLUSION

In this paper, we have proposed a novel way of exploring the problem of first-person authority and its exceptions by using the methods of experimental philosophy. Our approach is focused not on direct validation of philosophical theories or conclusions on that matter but on gathering empirical data about first-person authority understood as a special status enjoyed by self-ascriptions of mental states in everyday sociolinguistic practices. We observed that some assumptions about the strength and scope of first-person authority are unclear or controversial and therefore require empirical verification. We designed an experimental study investigating laypeople's reactions to scenarios in which self-ascriptions of mental states were overly questioned by a third-person observer. We have shown a difference in the strength of first-person authority between ascription of phenomenal and intentional states. We have also found that the closeness of the relationship between speakers may influence our acceptance of self-ascriptions being questioned, but only in the case of ascriptions of phenomenal mental states. Both these results may be used as an argument in favor of pluralist accounts of self-knowledge. Moreover, this result facilitates further research concerning other pragmatic aspects of first-person authority. Although our results on the matter of the status of disagreement concerning mental states are not conclusive, we have suggested possible directions for further research.

We believe that this approach and the results we present here open up promising possibilities for cooperation between the philosophy of mind and experimental philosophy.

APPENDIX A. STORIES

Groups 1 & 9: Suppose Tom is talking to his very close friend, Ben. Tom sincerely says, "I believe it's going to rain tomorrow." Ben disagrees, "No, you do not."

Groups 2 & 10: Suppose Tom is talking to his very close friend, Ben. Tom sincerely says, "I believe it's going to be sunny tomorrow." Ben disagrees, "No, you do not."

Groups 3 & 11: Suppose Tom is talking to his very close friend, Ben. Tom sincerely says, "I feel sad." Ben disagrees, "No, you do not."

Groups 4 & 12: Suppose Tom is talking to his very close friend, Ben. Tom sincerely says, "I feel happy." Ben disagrees, "No, you do not."

Groups 5 & 13: Suppose Tom is talking to his new colleague, Ben. Tom sincerely says, “I believe it’s going to rain tomorrow.” Ben disagrees, “No, you do not.”

Groups 6 & 14: Suppose Tom is talking to his new colleague, Ben. Tom sincerely says, “I believe it’s going to be sunny tomorrow.” Ben disagrees, “No, you do not.”

Groups 7 & 15: Suppose Tom is talking to his new colleague, Ben. Tom sincerely says, “I feel sad.” Ben disagrees, “No, you do not.”

Groups 8 & 16: Suppose Tom is talking to his new colleague, Ben. Tom sincerely says, “I feel happy.” Ben disagrees, “No, you do not.”

APPENDIX B. QUESTIONS

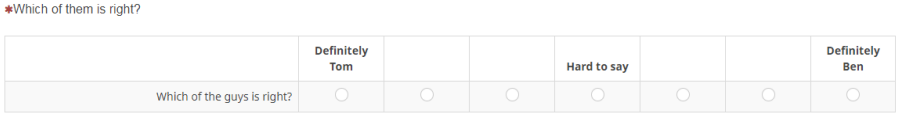
Participants in groups 1-8 were presented with the following question:


*Which of them is right?

	Definitely Ben			Hard to say			Definitely Tom
Which of the guys is right?	○	○	○	○	○	○	○

Which of them is right?

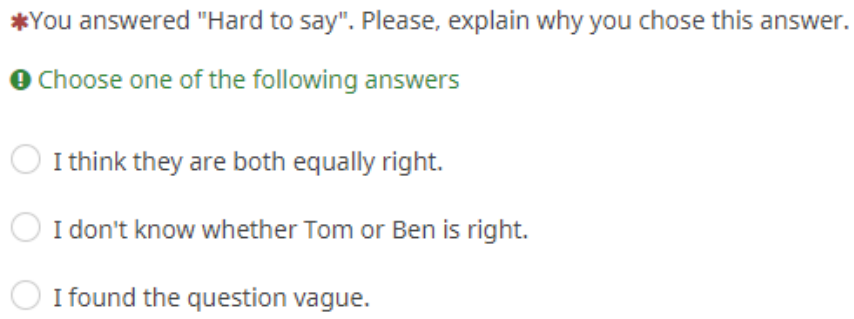
Participants in groups 9-16 were presented with the following question:


*Which of them is right?

	Definitely Tom			Hard to say			Definitely Ben
Which of the guys is right?	○	○	○	○	○	○	○

Which of them is right?

APPENDIX C: SUPPLEMENTARY QUESTION PRESENTED TO THE PARTICIPANTS WHO RESPONDED "HARD TO SAY"



APPENDIX D: FAULTLESS DISAGREEMENT QUESTION

For groups 3, 4, 7, 8, 11, 12, 15, and 16 was:

> *Is it possible that, despite their disagreement, neither of them is wrong about Tom's feelings?*

and for other groups:

> *Is it possible that, despite their disagreement, neither of them is wrong about Tom's beliefs?*

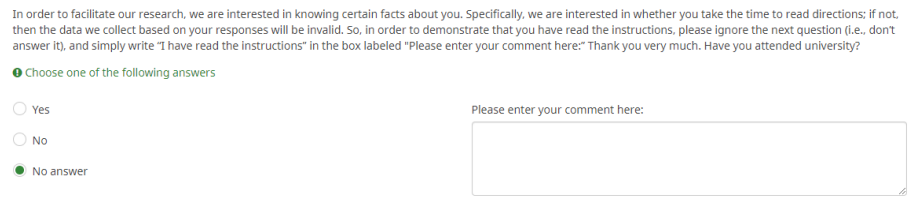
Possible answers to this question were "yes" and "no."

APPENDIX E: META-QUESTIONS

1. Was it difficult for you to understand the story?
2. Did you feel what Ben says in the story is strange?
3. Did you feel what Tom says in the story is strange?
4. Did you find the question "*Which of them is right?*" difficult?
5. Did you find the question "*Is it possible that, despite their disagreement, neither of them is wrong about Tom's beliefs?*" difficult?

For each of the questions, the possible answers were "yes" and "no". The answer was obligatory.

APPENDIX F: THE ATTENTION CHECK



BIBLIOGRAPHY

Bar-On D. (2004), *Speaking My Mind: Expression and Self-Knowledge*, Oxford: Oxford University Press. https://doi.org/10.1093/0199276285.001.0001

Boyle M. (2009), "Two Kinds of Self-Knowledge," *Philosophy and Phenomenological Research* 78(1), 133-164. https://doi.org/10.1111/j.1933-1592.2008.00235.x

Burge T. (1988), "Individualism and Self-Knowledge," *The Journal of Philosophy* 85(11), 649-663. https://doi.org/10.5840/jphil1988851112

Carruthers P. (2011), *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199596195.001.0001

Chalmers D. (2018), "The Meta-problem of Consciousness," *Journal of Consciousness Studies* 25(9-10), 6-61.

Coliva A. (2016), *The Varieties of Self-Knowledge*, London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-32613-3

Davidson D. (1984), "First Person Authority," *Dialectica* 38(2-3), 101-111. https://doi.org/10.1111/j.1746-8361.1984.tb01238.x

Finkelstein D. H. (2003), *Expression and the Inner*, Cambridge, MA: Harvard University Press.

Gertler B. (2001), "Introspecting Phenomenal States," *Philosophy and Phenomenal Research* 63(2), 305-328.

Goldman A. (2006), *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford: Oxford University Press. https://doi.org/10.1093/0195138929.001.0001

Jongepier F., Strijbos D. (2015), "Introduction: Self-Knowledge in Perspective," *Philosophical Explorations* 18(2), 123-133. https://doi.org/10.1080/13869795.2015.1032335

Kölbel M. (2004), "Faultless Disagreement," *Proceedings of the Aristotelian Society* 104(1), 53-73. https://doi.org/10.1111/j.0066-7373.2004.00081.x

MacFarlane J. (2014), *Assessment Sensitivity: Relative Truth and its Applications*, Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682751.001.0001

Moran R. (2001), *Authority and Estrangement: An Essay on Self-Knowledge*, Princeton: Princeton University Press.

Nichols S., Stich S. P. (2003), *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford: Clarendon Press — Oxford University Press. https://doi.org/10.1093/0198236107.001.0001

Rodriguez A. G. (2012), "How to Be an Expressivist about Avowals Today," *Nordic Wittgenstein Review* 1(3): 81-102.

Russell B. (1912), *Problems of Psychology*, New York: Henry Holt & Co.

Schwengerer L. (2019), "Beliefs over Avowals: Setting Up the Discourse on Self-Knowledge," *Episteme* 18(1), 66-81. https://doi.org/10.1017/epi.2018.56

Schwitzgebel E. (2008), "The Unreliability of Naive Introspection," *Philosophical Review* 117(2), 245-273. http://doi.org/10.1215/00318108-2007-037

Schwitzgebel E. (2012), *Introspection, What?* [w:] *Introspection and Consciousness*, D. Smithies, D. Stoljar (eds.), New York: Oxford University Press, 29-47. https://doi.org/10.1093/acprof:oso/9780199744794.003.0001

Stich S., Tobia K. P. (2016), *Experimental Philosophy and the Philosophical Tradition* [w:] *A Companion to Experimental Philosophy*, J. Sytsma, W. Buckwalter (eds.). https://doi.org/10.1002/9781118661666.ch1

Weinberg J., Alexander J. (2014), *The Challenge of Sticking with Intuitions Through Thick and Thin* [in:] *Intuitions*, A. Booth, D. Rowbottom (eds.), Oxford: Oxford University Press, 187-212. https://doi.org/10.1093/acprof:oso/9780199609192.003.0011

Wright C. (1998), *Self-Knowledge: The Wittgensteinian Legacy* [w:] *Knowing Our Own Minds*, C. Wright, B. C. Smith, C. Macdonald (eds.), Oxford: Oxford University Press, 101-122. https://doi.org/10.1017/S135824610000432X