DISSERTATIONES
PHILOSOPHICAE
UNIVERSITATIS
TARTUENSIS
26

NIKOLAI SHURAKOV

Experimental Contribution to Contemporary Ordinary Language Philosophy

DISSERTATIONES PHILOSOPHICAE UNIVERSITATIS TARTUENSIS
26

DISSERTATIONES PHILOSOPHICAE UNIVERSITATIS TARTUENSIS
26

NIKOLAI SHURAKOV

Experimental Contribution to Contemporary Ordinary Language Philosophy

UNIVERSITY OF TARTU
Press
1632

Faculty of Arts and Humanities, Institute of Philosophy and Semiotics

The Council of the Institute of Philosophy and Semiotics of the University of Tartu has accepted this dissertation for the defense for the degree of Doctor of Philosophy (in Philosophy) on September 8, 2025.

Supervisors: Associate Professor Alexander Stewart Davies and
Researcher Alex Wiegmann (University of Granada, Spain)

Opponent: Professor Vilius Dranseika, PhD, Jagiellonian University, Poland

Defense: The thesis will be defended at the University of Tartu, Estonia, on October 27, 2025, at 14.15 p.m. online (via two-way audiovisual equipment).

The research was supported by the Centre of Excellence in Estonian Studies (European Union, European Regional Development Fund); Doctoral Studies and Internationalisation Programme DoRa (European Union, European Social Fund through the Archimedes Foundation); the project "The Role of Metacontexts in Stabilizing Text Content: a case study of Estonia's Road Conditions Requirements Ordinance", PRG2151 (Estonian Ministry of Education and Research); a short-term grant by DAAD *(Deutscher Akademischer Austauschdienst)*; the project "Experimental Philosophy and the Method of Cases (EXTRA)" by DFG (*Deutsche Forschungsgemeinschaft*, grant number: 391304769); and the Graduate School of Linguistics, Philosophy and Semiotics (European Union, European Social Fund).

ISSN 1406-9520 (print)
ISBN 978-9908-57-009-9 (print)
ISSN 2806-2515 (pdf)
ISBN 978-9908-57-010-5 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF PUBLICATIONS

1. Shurakov N. (2025). The Stakes Effect: New Evidence from a Retraction-Based Experimental Design. *Episteme* 1–21. https://doi.org/10.1017/epi.2025.10060
2. Shurakov, N., & Wiegmann, A. (2025). The role of stakes in lying: An empirical investigation of the robustness of the folk concept of lying. *Philosophical Psychology*. https://doi.org/10.1080/09515089.2025.2506544
3. Shurakov, N., Kolbe, M., Engelmann, N., & Wiegmann, A. (forthcoming). GPT-4's alignment with human lie and falsity attribution in cases of deceptive implicatures. In V. C. Müller, A. R. Dewey, L. Dung, & G. Löhr (Eds.), *Philosophy of Artificial Intelligence: The State of the Art*. Synthese Library. Berlin: SpringerNature.
4. Davies, A., Rahuoja, F. G., & Shurakov, N. (2025). Abhorrent Slurs and Laudable Pejoratives: an Estonian Case Study of the Space Between. *Topoi*. https://doi.org/10.1007/s11245-025-10163-x

AUTHOR'S CONTRIBUTION

1. The article is the independent work of the author.
2. The author was involved in all stages of the research, including conceptualizing the initial idea, designing the experiments, and analyzing the results. Data analysis was conducted independently by both authors to verify the results. The author was responsible for writing the manuscript and implementing subsequent revisions.
3. The author was responsible for writing the manuscript and implementing subsequent revisions. Additionally, the author primarily wrote a Python script to interact with GPT-4 via API. The visualizations of the results were jointly produced by Kolbe and the author.
4. The author led the parts of the project that were conducted in Russian, interviewed participants, transcribed the interviews, and translated them into English. Also, the author contributed to the pre-interview stages, brainstorming interview questions and revising the original manuscript.

INTRODUCTION

This PhD dissertation brings together four articles that explore distinct yet interconnected topics within contemporary ordinary language philosophy, employing diverse empirical methods.

KNOWLEDGE: "The Stakes Effect: New Evidence from a Retraction-Based Experimental Design" investigates how practical stakes influence knowledge attributions. The experiments reveal a robust "stakes effect," where participants are significantly more likely to retract a knowledge claim after realizing the stakes are high. This holds not just for Dinges & Zakkou's (2021) experiments (where the retraction-based experimental design was introduced) but also for the direct replication, a third-person version of the Bank case, and for an experiment that implemented a modification addressing a couple of the shortcomings of Dinges and Zakkou's design. In this paper, I argue that these findings challenge the traditional view of knowledge – classic invariantism, and constitute some challenge to subject-sensitive invariantism. However, in its current form, the retraction-based design cannot distinguish between explanations offered by epistemic contextualism and relativism.

LYING: "The Role of Stakes in Lying: An Empirical Investigation of the Robustness of the Folk Concept of Lying" focuses on the concept of lying, examining whether stakes impact people's judgments about whether a statement is classified as a lie. At least some articles in the debates over the context-sensitivity of knowledge attributions have found an effect of stakes. This paper explores whether a similar effect exists for lying attributions. After obtaining a null result for the first experiment that tests low vs high stakes scenario pairs, we turn our attention to the difference between deontic statuses, i.e. cases where it is forbidden or obligatory to lie (we argue that this is a legitimate move, as stakes are often conceptualized as leaving a protagonist in a worse situation)[1]. Although we observed a significant effect of stakes in our Experiment 2, a more controlled version of it, i.e., Experiment 3, showed that the folk concept of lying is not significantly affected by either stakes or deontic status. This suggests that folk lying attributions are robust and not very sensitive to what is at stake.

GPT: "GPT-4's alignment with human lie and falsity attribution in cases of deceptive implicatures," explores, as its title suggests, the correlation between human concepts of lying and falsity and those of the large language model. It uses a cross-cultural human dataset (Wiegmann et al., 2024) as a benchmark to evaluate GPT-4's performance across six languages, again focusing on deceptive implicatures. The results indicate that while GPT-4's judgments strongly correlate with human judgments, GPT-4 tends to attribute lying more frequently than humans and falsity less frequently. This research has implications for the potential use of AI in identifying and correcting misinformation.

[1] I use 'we' for articles written with other people, occasionally switching to 'I' when I need to refer to my contributions

> PEJORATIVE: "Abhorrent Slurs and Laudable Pejoratives: An Estonian Case Study of the Space Between" is an empirically informed discussion of the ethical dimensions of pejorative language. It presents a case study of the Estonian pejorative 'tibla,' analyzing its use within the Estonian context. The study examines both the potential for laudable uses of the pejorative and the critiques to which it is susceptible, specifically when used as an ethnic-cognitive-behavioural pejorative. It explores how such a pejorative can be used to highlight existing injustices and challenge oppressive power structures, nonetheless remaining open to misuse (we demonstrate this with a brief case study application).

These four articles explore various conversational contexts and how these different contexts influence the things we do with words. All articles use empirical methods to capture how ordinary speakers use the target expressions. In this way, the dissertation is a contribution to the intersection between experimental philosophy and contemporary ordinary language philosophy.

The primary purpose of this framing chapter is to situate the dissertation as a contribution to *contemporary* ordinary language philosophy (contemporary OLP), that is informed by empirical methods (principally, but not solely, by experiments). In the first section, I demonstrate that OLP, after a historic decline, re-emerged, at least to some extent, due to the embrace of empirical methods. The rise of experimental philosophy (X-Phi) made it possible for OLP to overcome some of the initial criticism addressed to the tradition. Sections 2–5 provide a brief overview of each of the articles to extract statements put forward for the defence. KNOWLEDGE is an attempt to explore the alternative retraction-based experimental design when the two dominant methods struggle to move the existing debates forward. LYING employs a methodology similar to that in the debates over context-sensitivity of 'know', but applies it to investigate the folk concept of lying. GPT explores the possibility of using a large language model in the same way as humans participate in X-Phi studies. PEJORATIVE captures both laudable uses and critiques when the target expression is used. These sections distil from the individual articles the methods and the main findings presented for the defence.

Overall, the dissertation makes a significant contribution to contemporary ordinary language philosophy by using a range of empirical methods. I conclude with a short overview of how each of the articles falls into the intersection of contemporary OLP and experimental philosophy.

1. CONTEMPORARY ORDINARY LANGUAGE PHILOSOPHY AND X-PHI

This section situates the dissertation as a contribution to the intersection of experimental philosophy (X-Phi) and contemporary OLP. First, I briefly recap the traditional historical trajectory of OLP, which led the tradition to its perceived demise in the 1970s. One of the problems that has contributed to the original decline of OLP is its inability to verify claims about ordinary usage. OL philosophers used to rely on their intuitions, which did not always align even amongst OL philosophers themselves. Here, I argue that experimental philosophy offers contemporary OLP a means to overcome this limitation. By empirically investigating folk concepts, X-Phi enables philosophers to test ordinary use claims in a systematic and replicable way. But even if we can now more reliably track ordinary usage, a further question remains: *should* philosophers care about ordinary language at all? After all, the folk concept of time, for instance, plays little to no role in contemporary physics. Might the same be true in philosophy? To address this worry, I turn to an argument by Cappelen & McKeever (2022), who defend the relevance of ordinary concepts to many philosophical debates.

Taken together, these points motivate the research task addressed by the dissertation. It aims to make a contribution to contemporary OLP: one that appreciates the need for methodological rigour in investigating significant ordinary concepts. By borrowing or rather embracing empirical methods from X-Phi, contemporary OLP becomes well-positioned to advance a wide range of debates: from the nature of knowledge to AI. The dissertation contributes to this research programme by exploring several topics at the intersection of OLP and X-Phi, showcasing various experimental tools.[2]

1.1 On OLP

Illustrations of OLP are traditionally taken from the works of Austin, Ryle, Strawson and the later Wittgenstein. A characteristic feature of OLP is the special attention it pays to the *use* of relevant expressions:

> 132. We want to establish an order in our knowledge of *the use of language*: an order with a particular end in view; one out of many possible orders; not the order. To this end we shall constantly be giving prominence to distinctions which our ordinary forms of language easily make us overlook. (Wittgenstein, 1967, p. 52; emphasis added)

[2] This section does not aim to provide a comprehensive historical overview of the development of OLP and X-Phi. Instead, it constructs a framework within which the articles in the dissertation can be understood as collectively contributing to the same programme within philosophy.

However, after an initial period of quick growth, the tradition seems to have gone almost extinct, at least this is what "the standard story" says:

> There is a standard story told in histories of analytic philosophy about the rise and fall of ordinary language philosophy: it was a widespread, if not dominant, approach to philosophy in Great Britain in the aftermath of World War II up until the early 1960s, but with the development of systematic approaches to the study of language—formal semantic theories on one hand and Gricean pragmatics on the other—ordinary language philosophy more or less disappeared. (J. D. Porter & Hansen, 2023, p. 2; see also their list of references for this quote)

The method of OLP was primarily to analyze the situations in which a target expression is used. Here is how Hansen & Chemla (2015) describe the method of Austin:

> Austin's method of demonstrating fine distinctions between words or phrases with similar meanings involves constructing a type of linguistic experiment, in which we are either asked to judge whether it is better to use one or the other of the words being investigated in a particular situation, or to judge whether a particular expression better fits one or the other of a pair of situations that differ only in limited ways from one another. (Hansen & Chemla, 2015, pp. 423–424)

OL philosophers relied on their intuitions about what they thought people would say in a particular situation. Some examples:

> In no case do we say that… (Austin, 1962, p. 10)

> in an explicit performative we say 'I approve' rather than 'I approve of' (Austin, 1962, p. 84)

> and we say 'by "he" I was referring to…' (Austin, 1962, p. 97)

The method worked well for Austin, who is "regarded as having an especially acute ear for fine distinctions of meaning overlooked by other philosophers" (Hansen & Chemla, 2015, p. 422). Many did share Austin's intuitions regarding what we say. However, reliance on the "acute ear" of a particular philosopher also raises concerns about the arbitrariness of interpretation. How do we confirm that Austin is right or wrong about the ordinary use of a certain expression? Can we decide that the expression is used *correctly* in a given situation? What should we do in case of disagreements regarding which expression better fits? These questions are along the lines of the critique of OLP raised by Mates (1958, p. 161), who asked "how one would go about verifying these statements" about ordinary language use.

Mates' criticism of OLP is twofold. Firstly, he wonders how to verify one's diagnosis of ordinary use. When one reports ordinary use of an expression (make an OL claim), one may simply be wrong about it, or one's supporting evidence may not be that reliable. Specifically, Mates thinks that "Ryle's assertion about

the ordinary use of the word "voluntary" is false, and, moreover, that it is based upon doubtful evidence which would be entirely insufficient even if it were sound" (Mates, 1958, p. 163). While OLP has begun to successfully collect data on the use of ordinary expressions, it has not established a particular robust methodology that could have verified OL claims. The second part of Mates' criticism is the inability of OLP to deal with disagreements that arise when two OL philosophers have different views about the ordinary use of the same expression. When A says that X is used in one way, but R claims that it is used differently, how is the disagreement to be resolved?

Here's an example. Austin and Ryle disagreed about the ordinary meaning of '(in)voluntarily'[3]. Ryle took the view that....

> In their most ordinary employment 'voluntary' and 'involuntary' are used, with a few minor elasticities, as adjectives applying to actions which ought not to be done. We discuss whether someone's action was voluntary or not only when the action seems to have been his fault. He is accused of making a noise, and the guilt is his, if the action was voluntary, like laughing; he has successfully excused himself, if he satisfies us that it was involuntary, like a sneeze. (Ryle, 1949/2009, p. 56)

Whereas Austin took the view that ...

> For example, take 'voluntarily' and 'involuntarily': We may join the army or make a gift voluntarily, we may hiccough or make a small gesture involuntarily. (Austin, 1957, p. 17)

In Ryle's account, 'voluntary' and 'involuntary' are applied to the set of actions which "ought not to be done". For example, one is not expected to make noise during an academic talk at a conference. They ought not to laugh or sneeze. One is blameworthy for intentionally, voluntarily laughing, but is excused for an accidental sneeze. Austin lists some actions which seem to fall into the set that '(in)voluntarily' can modify: joining the army, making gifts, hiccough, making small gestures. If we try to merge Ryle's and Austin's respective treatments of '(in)voluntarily', we realize that both cannot always both be simultaneously correct, otherwise one "ought not to" join the army, make gifts, hiccough, or make small gestures.

This contradiction raises two questions. First, how did each philosopher arrive at their view about the ordinary use of '(in)voluntarily'? Both seem to rely on intuitions backed up with relevant examples. Their individual intuitions may not be representative of how ordinary speakers actually use the target expressions. Even eminent philosophers can overlook some patterns present in ordinary use. Without method(s) that make their claims verifiable, both Austin and Ryle are

[3] There are reasons to think Ryle and Austin have never actually been in a disagreement (See Zahorec et al., 2023, pp. 123–124). At least one is concerned with adjectives and the other is with adverbs.

essentially asking us to trust their individual assessments of what constitutes the ordinary meaning of '(in)voluntarily'. This reliance on unverified intuitions appears troubling when it leads to contradictory claims about the same target expression, as in the present case.

Second, how should such disagreement be resolved? In the absence of a shared standard for evaluating competing intuitions, there is no clear path forward, so Mates wonders "if agreement about usage cannot be reached within so restricted a sample as the class of Oxford Professors of Philosophy, what are the prospects when the sample is enlarged?" (Mates, 1958, p. 165). We simply have no way to resolve such disagreements if we don't have robust and reliable methods which could warrant the verification of OL claims.

In sum, Mates' critique consists of two interconnected points: we need a method for verifying claims about ordinary use (more reliable than the intuitions of particular philosophers), the lack of which makes disagreements which inevitably arise unresolvable. Next, I will examine two possible responses to this critique. The first states that verification is unnecessary. While Mates may be correct that OL claims lack verifiability, they do not require verification. The second approach is a suggestion to embrace empirical methodologies, attempting to validate ordinary language claims through evidence-gathering procedures similar to those employed in related disciplines such as linguistics or psychology.

The first response to Mates is to argue that no verification is needed for OL claims. It is a matter of fact that Austin and Ryle are highly competent native speakers of English. Cavell (1958) proposed that when competent speakers make claims like "we say X but not Y," they do not need to gather evidence:

> we must bear in mind the fact that these statements — statements that something is said in English — are being made by native speakers of English. Such speakers do not, in *general*, need evidence for what is said in the language; they are the source of such evidence. It is from them that the descriptive linguist takes the corpus of utterances on the basis of which he will construct a grammar of that language. (Cavell 1958, 174–175; emphasis original)

By virtue of being competent speakers of English, both Ryle and Austin are entitled to make OL claims. No verification is needed. Disagreements between competent speakers turn out to be faultless: no one is wrong when it comes to disagreeing with "we say X but not Y". Such a Cavell-style response to Mates' challenge faces a straightforward counterexample.[4]

Let us imagine a speaker of English, Jack, who consistently misuses a particular expression—he uses E instead of Y. Suppose that Jack even insists that this is how E is ordinarily used and does not accept any correction. As a competent speaker, he might claim the authority to assert the OL claim regarding E, much like Austin or Ryle did regarding '(in)voluntarily'. Jack claims: "we say

[4] Cavell (1958) has more to offer explaining the Ryle/Austin disagreement, but what I am trying to demonstrate here is that 'no verification is needed' is not an optimal response to the Mates' challenge.

E but not Y". Treating verification of such claims as unnecessary vindicates (almost) any linguistic usage, no matter how idiosyncratic, by competent speakers. OLP then risks being committed to the conclusion that no usage claim can be meaningfully challenged or corrected (even in cases like Jack's). "Ordinary" becomes reduced to "whatever a competent speaker says." If two competent speakers assert conflicting OL claims, and no method exists for assessing whose account is more accurate (might be both, might be none), OLP might well struggle to deal with competing interpretations. To avoid such cases, OLP should not treat verification as unnecessary *pace* Cavell.

The second response available to OLP is to accept that Mates has pointed to an important issue in its methodology. There is no guarantee that any given OL philosopher will differ significantly from the hypothetical Jack described earlier. Traditional OL philosophers often generalized their own intuitions to what *we* say (e.g., see examples from Austin above), where 'we' tries to denote the population of all competent speakers. This latter sample differs crucially from the "the class of Oxford Professors of Philosophy": the sample of speakers of English is more diverse in race, sex, education, social class and so on. Sometimes it is simply the case that "ordinary speakers do not say what philosophers say they will say (Schaffer & Knobe, 2012, p. 703). It is largely a mystery how Austin was coming up with his definitions or sharp distinctions in what ordinary expressions mean. Even geniuses make mistakes. Establishing some robust procedures that lead to OL claims could help OLP point out such occasional mistakes and resist hypothetical Jacks. Adopting methods that make OL claims verifiable is another way to overcome Mates' critique.

Such a move has already been proposed by Katz and Fodor (1963), who argue that OL claims require empirical justification, especially once two native speakers disagree:

> That Cavell's position blocks an adequate understanding of ordinary language philosophy follows from the fact that the Oxford philosopher, when he discusses the use of words, *is* pursuing an empirical investigation, and *is not* uncovering truths of transcendental logic. (Fodor & Katz, 1963, p. 16; emphasis original)

Obviously, empirical investigation requires empirical methods. If philosophical claims about ordinary meanings are, in principle, claims about linguistic behaviour, then these claims are subject to empirical scrutiny. What should we expect from empirical methods to be verifiable? Ideally, they should make the process by which OL claims are obtained transparent. For example, data collection should be conducted systematically rather than through the cherry-picking of examples that occur to an OL philosopher. Generalizations must be justified through established procedures (e.g., by using inferential statistics). We should be able to see the hypotheses of a study and how the findings are reached. In this way, other researchers can attempt to replicate, and thereby *verify*, the findings. When disagreements about usage arise, one can evaluate the empirical evidence and the procedures through which it was gathered to determine which

position enjoys stronger support. It seems that adopting such methods (perhaps borrowing something from sister disciplines like psychology) could have been beneficial to OLP.

Unfortunately, Katz & Fodor's proposal was not adopted by OL philosophers of the 20th century, so the problem of having methods of gathering the data of ordinary use that are verifiable has remained. Perhaps, this methodological shortcoming contributed to the decline of the tradition. However, the good news is that empirical methods were eventually embraced by philosophers several decades later, giving rise to experimental philosophy.

1.2 Experimental Philosophy as an Ally to OLP

The previous section discussed Mates' challenge to OLP emphasizing the need for verifiable methods. In this section, I explore how experimental philosophy (X-Phi) provides a promising response to this critique. I begin by outlining the emergence and methodological development of X-Phi. I then show how experimental tools can be used to support OLP, allowing philosophers to empirically investigate OL claims.

The emergence of experimental philosophy in the early 2000s allowed philosophers to "go into the wild," collecting data on how concepts are deployed by ordinary speakers. This methodological innovation contributed to a revival of OLP, sometimes referred to as the "new wave" of OLP (Hansen, 2014; J. D. Porter & Hansen, 2023). New-wave, contemporary OLP continues to focus on language use but is now enhanced by a variety of experimental tools that can systematically investigate relevant distinctions. In what follows, I provide a sketch of the rise of experimental philosophy, adding empirical methods to philosophers' toolbox. I argue that experimental philosophy became a great ally to OLP, helping it to overcome Mates' critique.

The analytic philosophy of the second half of the 20th century is full of famous cases: the Chinese Room, Gettier' cases, Twin Earth, Mary's Room, and many others. Using this method of cases, "philosophers routinely appeal to intuitive judgments about hypothetical thought experiment cases, which are then used to support or refute philosophical theories that make a prediction about how these judgments should turn out" (Horvath, 2023, p. 72). X-Phi emerged in the early 2000s, partially as a response to this reliance on intuitions in analytic philosophy (Knobe & Nichols, 2008). Proponents of X-Phi argued that empirical studies should inform philosophical theories of how people actually think and talk rather than merely relying on the intuitions of philosophers (even Oxford philosophers doing OLP). The first X-Phi findings indicated that there are cross-cultural differences in intuitions about knowledge (Weinberg et al., 2001) and reference (Machery et al., 2004), suggesting that philosophical intuitions may not be as universal as previously assumed. The common method used in these (and many other) papers was to survey participants about their intuitions regarding relevant cases.

The bibliometric analysis conducted by Li & Zhu (2023) of X-Phi papers demonstrates that "X-Phi has undergone roughly four developmental stages over the past two decades, namely, the initiation period (2000–2005), the development period (2006–2010), the expansion period (2011–2015), and the plateau period (2016–2020)" (Li & Zhu, 2023, p. 47). During its development, X-Phi has expanded in both its methods and ambitions, leading to narrow and broad definitions of what X-Phi is. In the narrow sense, X-Phi is still understood as primarily dealing with intuitions regarding cases:

> In a narrower sense, 'experimental philosophy' refers to a more specific kind of philosophically motivated experimental inquiry, in which verdicts on hypothetical cases relevant to some philosophical question are elicited from significant numbers of subjects. (Williamson, 2016, p. 22)

The broader definition is more inclusive in two ways. Firstly, on this definition, X-Phi aims to research phenomena beyond those described by, or argued for with, the method of cases. And secondly, it adopts a wider set of methods, stepping beyond vignette-based experiments and questionnaires, which were often used in the first X-Phi articles. Sytsma et al. (2023) provide a definition of X-Phi that includes both of these dimensions. They describe X-Phi simply as philosophy "with a little something extra," with that extra component mostly being the fact that "experimental philosophers *do* empirical science as part of *doing* philosophy" (Sytsma et al., 2023, p. 10; emphasis in original). In this dissertation, I sign up for this broad definition of X-Phi. Each of the contributing articles is also an X-Phi article according to this definition (although KNOWLEDGE and LYING also count as X-Phi articles in the narrow sense).

It is not a simple coincidence that the seminal X-Phi work Knobe (2003) is titled "Intentional action and side effects in ordinary language". X-Phi becomes a perfect ally for contemporary OLP because Mates' critique of OLP may now be addressed. X-Phi studies can test claims about ordinary usage that OLP philosophers previously made based on intuition alone or a very limited sample. Of course, not all contemporary OLP philosophers are experimental philosophers (nor should they be), but the overlap between these two areas of philosophy is evident. Such overlap is acknowledged in the literature. In particular, Mizumoto (2023, p. 31) "show[s] that the contemporary meta-philosophical debates over X-Phi are in fact a revival of the debates in the heyday of OLP. This will then also indicate a parallel between Wittgenstein's negative program, called quietism, and the negative program of X-Phi, especially Stich's." Hansen (2014) argues that "contemporary versions of ordinary language philosophy are alive and flourishing but going by various aliases—in particular (some versions of) 'contextualism' and (some versions of) 'experimental philosophy'" (Hansen, 2014, p. 556). So, what does X-Phi offer to contemporary OLP?

X-Phi employs empirical methods of psychology or cognitive science (Knobe, 2007, 2016), like controlled experiments, surveys, and qualitative interviews, and this list constantly expands. It often collects and evaluates data from large and

demographically diverse samples (e.g., Rose et al., 2019). Cova et al. (2021, p. 10) "found that X-Phi studies—as represented in our sample—successfully replicated about 70% of the time". Such high replicability emphasises the overall reliability of evidence provided by experimental philosophers. In this way, X-Phi equips OL philosophers to empirically test OL claims, finding verifiable and replicable results. In doing so, X-Phi directly confronts Mates' critique: firstly, one can easily elaborate on the empirical method used on the way to an OL claim; and secondly, disagreements that once seemed irresolvable because they stemmed from conflicting intuitions can now be addressed through a systematic survey of the existing evidence. When philosophers disagree about how a term is ordinarily used, experimental methods provide a shared framework for investigation, establishing a common ground from which we can evaluate competing OL claims. By turning to empirical evidence, X-Phi helps transform OLP from a tradition reliant on the philosopher's "acute ear" into a practice grounded in statistics and empirical methods.

Let us return to the *apparent* disagreement between Austin and Ryle concerning the ordinary use of '(in)voluntarily' to see how X-Phi can deal with it. Drawing on data from the British National Corpus (BNC), Zahorec et al. (2023) conducted an X-Phi study, combining corpus data and clustering techniques, to evaluate whose characterization (Austin's or Ryle's) better aligns with actual usage. Crucially, their analysis revealed that although their analyses align with Austin's and Ryle's OL claims (mostly with Austin's), both philosophers overlooked what Zahorec et al. call a *physiological* use of '(in)voluntarily'. Such uses center "on whether an action or movement is in some way automatic rather than under conscious control" (Zahorec et al., 2023, p. 143). As Zahorec et al. conclude, "we have vindicated Mates' general worry... as we have seen how both Ryle and Austin overlook some types of ordinary uses of these expressions and how that leads to incomplete general observations about the features of ordinary language" (Zahorec et al., 2023, p. 146). This study substantiates Mates' general concern: it shows how easily certain use cases can go unnoticed by even the best OL philosophers. Austin's treatment of '(in)voluntarily' while adequate and supported by instances from the BNC, nonetheless overlooks an important part of the ordinary meaning of the target expression. This X-Phi study illustrates how useful empirical methods might be in providing an adequate picture of how ordinary expression is used: instead of relying on the intuitions of Austin and Ryle, Zahorec and colleagues looked at the huge amount of data from samples of written and spoken (British) English. They can hardly be accused of cherry-picking the examples; the method they used is transparent, and anyone who can use a corpus can re-run the study to verify their findings. The broader lesson, then, is methodological. Experimental methods appear not to be supplemental but essential for a significant part of contemporary OLP.

1.3 Cappelen and McKeever's Defence of OLP

The two previous subsections demonstrate that X-Phi helps OLP to overcome Mates' critique, so contemporary OLP, enhanced with empirical methods, is a viable program. It *can* be done. But *should* it be done? Why is the study of ordinary language usage important to philosophy? This section provides a standalone motivation for OLP. It briefly restates the argument for philosophers to pay attention to ordinary concepts. This argument not only justifies the research done in each particular article included in this dissertation but also motivates research within the programme of (contemporary) OLP as a whole.

Cappelen & McKeever (2022) argue that some philosophical investigations are OL guided:

> OL guidance: An investigation is OL guided if it aims to investigate (and understand) those parts of the world that a particular ordinary language expression picks out and, in so doing, is in part guided by data about the relevant ordinary language terms (Cappelen & McKeever, 2022, p. 225)

They provide two motivations for OL guidance: CARE and ANCHORING. CARE posits that philosophers are concerned with, i.e. care about, the ordinary meaning of terms (Cappelen & McKeever, 2022, pp. 227–228). We can see this by understanding why epistemologists care about the use of 'know'. There are arguments for scepticism which involve a radical sceptical hypothesis (like the brain-in-a-vat hypothesis) and a closure principle. The principle of epistemic closure states that if a person knows that *p*, and knows that *p* entails another proposition *q*, then that person also knows that *q*. According to the radical sceptical hypothesis, one cannot know whether one is a brain-in-a-vat, as the perceived experience is indistinguishable between being a brain-in-a-vat and being an ordinary human being. Combining closure and radical scepticism, one has to conclude that one does not know very ordinary things (e.g., one cannot know that one has hands). Importantly, scepticism does not rely on any special technical version of knowledge ($know^{high}$ or $know^{strong}$). We would probably not be bothered much if we were unable to reach $know^{high}$. The point of such scepticism is in undermining our mundane, ordinary knowledge expressed with the ordinary use of 'know':

> It would probably be quite clear to us that the skeptic was doing something fairly new and different when she started using 'know(s)' in the 'high octane' sense, and it would probably be pretty obvious she was 'changing the subject' on us, and the arguments would not seem to be a threat to our ordinary knowledge. (DeRose, 2009, p. 43)

For this reason, epistemic contextualism does not introduce $know^{high}$, i.e. a special concept operated by a sceptic (or employed only in the high stakes situations), whilst leaving a separate concept $know_{low}$ for ordinary use. It instead draws attention to the different epistemic standards governing various conversational

contexts. In this way, a proponent of epistemic contextualism cares about the *ordinary* use of 'know'.

The second motivation for OL guidance is ANCHORING: "The meanings and denotations of ordinary language expression serve as anchoring points for conceptual vicinities ... The process of vicinity creation starts with an understanding of T" (Cappelen & McKeever, 2022, p. 230, where T refers to an ordinary language term). Ordinary expressions often set the terms of philosophical discussions. For example, the folk concept of lying not only deserves to be investigated as such but also should serve as a starting point (i.e. anchor) for revisionary projects:

> Lying is a concept used in everyday language, and moral questions about lying arise in people's everyday experience. There are no compelling reasons to revise or reject the ordinary language concept of lying—at least the burden of proof rests with those who would revise or reject it. (Carson, 2010, p. 33)

Motivated by CARE and ANCHORING, many parts of philosophy are, indeed, OL guided. Philosophers do care about the ordinary meaning of "know" or use the folk concept of lying to anchor their theories of what lying is. These are not technical notions (although one is never deprived of introducing such), but instead ordinary terms that some philosophers deal in. However, there has always been a challenge in verifying or accurately capturing the content of ordinary language concepts. X-Phi provides rigorous empirical techniques to do this.

Let's recap. Firstly, we have seen that Mates' critique of OLP concerns the problem of verifying philosophers' OL claims. This challenges the very feasibility of ordinary language analysis. However, X-Phi furnishes tools capable of addressing such worries in a systematic and empirically grounded way. Secondly, there exists an independent motivation for doing OLP, namely, that some parts of philosophy are OL-guided. At least in these parts, empirically informed contemporary OLP has the potential to either initiate new or move the existing debates further. Therefore, the research task for this dissertation is to explore this potential and contribute to the revival of OLP.

With this background in place, let's turn to the content of this dissertation. The dissertation makes an experimental contribution to several ongoing debates within contemporary OLP. The following sections provide a brief overview of the debates each of the papers contributes to, describe the methods used and end with a specific statement presented for a defence.

2. KNOWLEDGE

To repeat, Hansen (2014) claims that "contemporary versions of ordinary language philosophy are alive and flourishing but going by various aliases—in particular (some versions of) 'contextualism' and (some versions of) 'experimental philosophy'" (Hansen, 2014, p. 556). The first paper engages in both contextualism and experimental philosophy: it experimentally tests the predictions made by epistemic contextualism. In what follows, I briefly introduce the contextualist prediction, then I sketch the experiments conducted in KNOWLEDGE testing such prediction and report the findings. I conclude the section by restating the main findings and outlining directions for future research.

2.1 Motivation

Epistemic contextualism is the view that the truth-conditions of knowledge attributions can shift depending on the context. The Bank case is often used to illustrate the position:

> Bank Case A. My wife and I are driving home on a Friday afternoon. We plan to stop at the bank on the way home to deposit our paychecks. But as we drive past the bank, we notice that the lines inside are very long, as they often are on Friday afternoons. Although we generally like to deposit our paychecks as soon as possible, it is not especially important in this case that they be deposited right away, so I suggest that we drive straight home and deposit our paychecks on Saturday morning. My wife says, "Maybe the bank won't be open tomorrow. Lots of banks are closed on Saturdays." I reply, "No, I know it'll be open. I was just there two weeks ago on Saturday. It's open until noon."
>
> Bank Case B. My wife and I drive past the bank on a Friday afternoon, as in Case A, and notice the long lines. I again suggest that we deposit our paychecks on Saturday morning, explaining that I was at the bank on Saturday morning only two weeks ago and discovered that it was open until noon. But in this case, we have just written a very large and very important check. If our paychecks are not deposited into our checking account before Monday morning, the important check we wrote will bounce, leaving us in a very bad situation. And, of course, the bank is not open on Sunday. My wife reminds me of these facts. She then says, "Banks do change their hours. Do you know the bank will be open tomorrow?" Remaining as confident as I was before that the bank will be open then, still, I reply, "Well, no. I'd better go in and make sure." (DeRose, 1992, p. 913)

In both cases, the evidence available to the protagonist is the same, but what is at stake is different. DeRose argues that both knowledge ascription in Case A and knowledge denial in Case B are true, so we face an apparent contradiction. DeRose (2009) also reports a conversation with his advisor—Rogers Albritton, who said:

> ‘Look, *if you’re going to do ordinary language philosophy*—and that’s what you’re doing here—you’d better do it right’
>
> Albritton never explained to me why the examples should be constructed so that what’s said is natural and appropriate beyond insisting that that’s how ordinary language philosophy should be done. (DeRose, 2009, p. 51; emphasis is mine)

What DeRose does is OLP: he attempts to demonstrate that ordinary speakers find both knowledge ascription in Case A and its denial in Case B true. In this way, contextualism posits a semantic thesis on context-sensitivity of ‘know’. DeRose (2005, p. 172) even argues that ordinary speakers’ intuitions provide evidence of “the best type” for *any* piece of language being context sensitive. Thus, empirical methods are an obvious choice for addressing whether these predictions regarding the Bank case are correct. My first article deals with the question: Do stakes influence knowledge attributions?

2.2 Description of Experiments

Many studies have attempted to establish whether contextualism (or its rivals) accurately predicts ordinary speakers’ judgements.[5] But the results, so far, have been inconclusive. There are two dominant experimental designs: evidence-fixed and evidence-seeking (see Section 1 of KNOWLEDGE for a detailed overview of the experimental literature). In brief, the evidence-fixed design provides participants with scenarios like Bank Case A or Bank Case B and asks them whether they agree with the knowledge attribution made or whether it is true. The evidence-seeking design explores the perceived amount of evidence protagonists possess by asking questions like “How many times do you think Peter has to proofread his paper before he knows that there are no typos?”(Pinillos, 2012, p. 199), expecting that more checks are needed when stakes are high. As I demonstrate in KNOWLEDGE, both experimental designs struggle to provide conclusive evidence for debates: some experiments employing the evidence-fixed design have detected stake effects, while others have not. Regarding the evidence-seeking design, concerns have been raised about whether this paradigm tests knowledge attributions per se or rather the evidential threshold for knowledge (see Porter et al., 2024). Might an alternative experimental design illuminate these issues?

KNOWLEDGE adopts an alternative design: the retraction-based experimental design introduced by Dinges & Zakkou (2021). Their design targets the *use* of knowledge attributions:

[5] This list includes (Buckwalter, 2010; Buckwalter & Schaffer, 2015; Feltz & Zarpentine, 2010; Francis et al., 2019; Francis & Beaman, 2023; Grindrod et al., 2019; Hansen & Chemla, 2013; May et al., 2010; Pinillos, 2012; Pinillos & Simpson, 2014; B. Porter et al., 2024; Rose et al., 2019; Schaffer & Knobe, 2012; Sripada & Stanley, 2012; Turri, 2017; Wu, 2023)

> In this design, participants are asked to read a setup of a situation that ends with a knowledge (self)ascription, e.g. after reading about having a paycheck to deposit and remembering being to the bank three weeks ago: "based on this, you respond: "I know the bank will be open tomorrow"."(D&Z, p. 736). The next part of the scenario is different, depending on which of three conditions a participant has been assigned to: NEUTRAL (no new information regarding the knowledge claim provided), STAKES (high stakes introduced), or EVIDENCE (evidence against *p* is offered). The crucial question then is whether participants stand by their claim that they know the bank will be open tomorrow. Choosing between 'I do' and 'I don't', they are allowed to retract the knowledge claim. The results are robust. The retraction rates for the Bank scenario are as follows: retraction in STAKES (48%) is significantly higher than in NEUTRAL (9.8%), and retraction in EVIDENCE (96.1%) is higher than in STAKES. A similar pattern holds for the Typo scenario (wherein a student needs to proofread a paper). The retraction rates were as follows: NEUTRAL 5.9%, STAKES 24%, EVIDENCE 58%. Across both scenarios, D&Z consistently found a medium to large effect of stakes. (KNOWLEDGE, p.8)

In KNOWLEDGE, I report the results of two experiments. Experiment 1 consists of a direct replication of Dinges & Zakkou's experiment that used the Bank scenario and tested whether their findings extend to third-person cases. Experiment 2 modified the initial design to increase the validity of their findings. All experiments are preregistered, and data and analysis files are openly shared[6].

Experiment 1 has the following structure. After providing informed consent, participants were randomly assigned to one of the conditions: 3x2 design, 3 story types (Neutral, Stakes and Evidence; between subjects) x 2 perspectives (first or third-person knowledge attributions; between subjects). Participants read a scenario and then were asked three questions: whether they stand by the knowledge ascribing claim, a confidence question (a 7-point Likert item ranging from "very unconfident" to "very confident") and an attention check. The first-person version of the Bank scenario was a direct replication of Experiment 2 from Dinges&Zakkou's article. The third-person version was obtained with minimal intervention to the original text.

Experiment 2 modified the design of the experiment by introducing a binary choice after the story setup. I gave participants a choice to either attribute knowledge or not; they could choose between I know vs I don't know that the Bank is open. As I argue in the article, this modification of the initial design improved the overall ecological validity of the results. Firstly, it made the act of retraction, which is a primary task in the experiment, more realistic: it is a deliberate choice of participants to endorse the knowledge claim. Secondly, it allowed the exclusion of the scenario sceptics—people who might have found the available evidence below the knowledge threshold. I employed only the first-person version of the Bank case in Experiment 2. This modification provided participants

[6] The project is available here: https://osf.io/tys3p/
The preregistrations: https://osf.io/5z6pu (Experiment 1), https://osf.io/ae3pu (Experiment 2)

with an opportunity to show how *they* use knowledge attributions: they made a choice whether to ascribe and later whether to stand by a knowledge attribution. In this way, the retraction-based design shares the spirit of OLP.

2.3 Main Findings and Future Research

Statement for the defence: Using the retraction-based experimental design, there is a new piece of evidence that stakes do influence knowledge attribution.

When participants are told that the stakes are high, they are more likely to retract the knowledge ascription made at a low-stakes setup. The retraction rates in Neutral (nothing important happens) conditions are lower than in Stakes (high stakes for a subject of knowledge are introduced), and both are lower than retractions in the Evidence condition. This pattern was consistently reproduced in several experiments. Firstly, it holds for Experiment 2 (Dinges & Zakkou, 2021) and its direct replication. Secondly, the same is true regarding the third-person version of the Bank case. Thirdly, the same pattern replicates in an experiment with a modified design. The findings of this series of experiments indicate that a significant effect of stakes might exist, and its size is medium to large. In KNOWLEDGE, I argue that the results challenge rivals of epistemic contextualism: classic invariantism and subject-sensitive invariantism, but they also do not distinguish between contextualism and relativism: both may interpret the results in their favour.

Besides, the retraction-based experimental design has been shown to be an attractive alternative to the dominant paradigms (evidence-fixed and evidence-seeking), capable of providing novel evidence to the debate.

The most notable limitation of the paper is the use of only one scenario—the Bank case. Thus, the question of generalizability remains. Taking into account the effect sizes (medium to large across all experiments) and the Dinges & Zakkou's experiment, which tested another scenario called 'Typo', I expect that the effect may still be generalized for other scenarios. Another obvious step for future research is to explore whether the findings hold for more third-person scenarios in a modified design. Finally, there is room for more questions in the retraction-based paradigm—one must examine the motivations for not standing by the claim, e.g. if one retracts because the knowledge attribution done in the setup is no longer true. In this way, the retraction-based design might be improved even further in order to distinguish between epistemic contextualism and relativism.

There are also more general directions in which the debate over the stakes-sensitivity of knowledge attributions could be taken. There is a considerable body of experimental research that has yet to receive adequate theoretical reflection. There are three experimental designs which have been used in attempts to address the same question, and the experimental results are still mixed. This experimental work could greatly benefit from more theoretical reflection on these mixed results (e.g., like Dinges (2023)).

The diversity of findings might also be addressed through further methodological improvements. Drawing from experiments in economics and psychology, Hansen (2020, sections 4.1–4.3) has already outlined the following means of improvement: varying actual stakes for participants (as is often done in economics), exploring real-world settings, making use of participants who are not aware that they are being observed, and incorporating conversational dynamics into studies. Each of these changes may contribute to more rigorous methodologies.

"There is continuity in the underlying debate about the relation between the ordinary use of the word "know" and the nature of knowledge from the 1940s through the 2000s" (Porter & Hansen, 2023, p. 225) . The first paper of the dissertation contributes to this continuity. It tests the prediction that our ordinary use of 'know' is stakes-sensitive and finds robust evidence that it is. It deals with the concept epistemologists care about (in accordance with CARE above). Importantly, the experimental design employed was focused on the *use* of knowledge attributions: whether participants find them acceptable after the setup and whether they retract them. Thus, KNOWLEDGE is positioned as an experimental paper at the intersection of contemporary OLP and X-Phi.

3. LYING

This article continues the research of the possible impact of stakes on concept use. It explores whether the folk concept of lying is stake-sensitive. In what follows, I motivate the research, overview the experiments conducted, and present the main findings.

3.1 Motivation

Lying has been a subject of philosophical inquiry throughout history, from Aristotle's discussion in *Nicomachean Ethics* and St. Augustine's *De Mendacio* to Kant's categorical imperative against falsehood in *Groundwork of the Metaphysics of Morals* and more contemporary analytical approaches. While this historical trajectory has produced plenty of sophisticated normative accounts, experimental philosophy contributes to these inquiries by providing novel evidence on the *folk* or *ordinary* concept of lying. The experimental turn in the philosophy of lying is grounded in the idea that a definition of lying should reflect (or at least, depart from) *how* ordinary people *use* this concept:

> Lying is a concept used in everyday language, and moral questions about lying arise in people's everyday experience. There are no compelling reasons to revise or reject the ordinary language concept of lying—at least the burden of proof rests with those who would revise or reject it. Therefore, consistency with ordinary language and people's linguistic intuitions about what does and does not count as a lie is a desideratum of any definition of lying. (Carson, 2010, p. 33)

In other words, Carson emphasizes the importance of engaging with the ordinary language concept of lying. Here, the folk concept of lying continues to serve as an anchoring point for any other theory of lying (recall the ANCHORING motivation for philosophers' interest in ordinary language). Thus, the article, by exploring a particular feature of the concept, contributes to contemporary OLP.

The current philosophical landscape presents a dichotomy between relatively restrictive definitions (Saul, 2012; Stokke, 2013) and more broad characterizations (e.g., Viebahn, 2021) of what constitutes lying. Notably, Stokke (2013, p. 49) defends a common ground approach: "S lies to X if and only if (L1) S says that p to X, and (L2) S proposes that p become common ground, and (L3) S believes that p is false." Viebahn (2021, p. 292) presents a broader assertion-based formulation: "A lies to B if and only if there is a proposition p such that: (L1) A asserts p to B, and (L2) A believes that p is false." None of these definitions accommodates the potential influence of stakes on lying attributions. The following provides three reasons to believe that, contrary to these definitions, stakes do influence the interpretation of lying attributions.

First, a stakes effect has been documented across some domains in philosophy. In epistemology, a substantial body of research demonstrates that stakes can

influence ordinary knowledge ascriptions (see Pinillos (2016), for an overview, and the footnote references in the previous section). And recall that KNOWLEDGE found a medium to large effect of stakes across both experiments. It is reasonable to hypothesize that the effect of stakes might extend to other phenomena, e.g., lying attributions. This is particularly relevant for proponents of knowledge-lying accounts (notable examples include the known-false account in (Holguín, 2021; Turri, 2016, pp. 34–36), which posit that S lies only if S knows that *p* is false. If 'know' is sensitive to stakes, then by extension, lying attributions must be similarly sensitive.

Second, high-stakes scenarios frequently involve significant moral considerations (DeRose's Bank Case B postulates that you may end up in a *very bad* situation). Moral considerations have been demonstrated to influence people's judgments. The "Knobe effect" (Knobe, 2003) demonstrates that when an action results in harm, individuals are more likely to attribute intentionality to the agent compared to when the action leads to a positive outcome. Furthermore, such an effect is argued to influence a range of other concepts, such as knowledge, happiness, value, act individuation, freedom, and naturalness (see Knobe (2010) for an overview). Although lying is generally considered an intrinsically bad form of speech, it seems plausible that lying attributions might be affected by the moral stakes involved in a given situation.

Third, a proper understanding of high-stakes lying has significant utility for other disciplines, including economics (Gneezy, 2005; Gneezy & Kajackaite, 2020; Gylfason et al., 2023; Hu & Ben-Ner, 2020) or behavioural psychology (Frank & Ekman, 1997; O'Sullivan et al., 2009; Wright Whelan et al., 2015). Crucially, many of these studies only *assume* the existence of 'high-stakes lying' and hypothesize that high stakes might trigger distinctive behaviours but it is unclear whether such type of lies even exists.

3.2 Description of Experiments

LYING consists of three experiments. All used a similar design. A quantitative experimental method seemed optimal for this research task, as we tried to test whether what is at stake affects lying attributions. We employed a traditional context-shifting experimental design (see (Hansen, 2012)): a protagonist utters the same target sentence in two contexts, C1 and C2, which are different with respect to what is at stake (broadly understood). The proposition is true according to a described state of affairs but generates a deceptive implicature. To illustrate, here is the *Peanuts* scenario employed in Experiment 1:

> George is cooking dinner for Frida, whom he does not really like. He knows that
> Variant A: Frida does not like the taste of peanuts,
> Variant B: Frida has a deadly peanut allergy,

> so he asks his flatmate Hans to sneak some peanuts into the meal while George and Frida are on the balcony. Frida, being rightly cautious, asks whether George has put any peanuts in the meal.
>
> George replies: "I didn't put any peanuts in the meal." (LYING, p. 1)

What George has said is true—it was Hans who put the peanuts into the meal, but George's utterance generates an implicature that there are no peanuts in the meal (that is false). We decided to test deceptive implicatures because they might be perceived as borderline instances of lying. Existing literature on deceptive implicatures (Reins et al., 2023; Weissman & Terkourafi, 2019; Wiegmann, 2023b, 2023a; Wiegmann et al., 2024) identified that there exists a significant variation in whether ordinary speakers interpret them as lies.

In each of the experiments, participants were first asked to morally evaluate the actions of the protagonist using a 7-point Likert item, which ranged from 'not bad at all' to 'very bad'. This was introduced to separate their moral evaluation of the protagonist's action from their assessment of lying. One can think that George's actions were bad, but he didn't lie. The second question was to directly assess whether they believed the protagonist had lied, with a simple 'yes' or 'no' response. An attention check was included at the end of each condition to ensure that participants paid attention to the task.

The first experiment that tested high vs low-stakes pairs of scenarios (e.g., Peanuts above) got a null result. We established that answers to the first question (morality) were not much different between the low vs high stakes conditions, so we increased the differences between the stories, operationalizing stakes as deontic status: forbidden vs obligatory to lie. As the stakes are conceptualized quite broadly in the literature, this move seemed legitimate. Moreover, while it is mostly forbidden to lie in our everyday contexts, cases where lying is expected from a speaker typically involve high stakes: Kantian Murderer at the Door is a clear example. Our Experiments 2 and 3 explored this possibility. In order to increase potential generalizability, each of the three experiments utilized two scenarios (within-subject).

3.3 Main Findings and Future Research

Statement for the defence: The stakes (broadly construed) do not significantly impact lying attributions.

In a series of experiments, we found that the folk concept of lying is robust enough and is not moved by stakes (or stakes operationalized as deontic status). Although what we got in Experiments 1 and 3 are null results, the power of our experiments suggests that even if stakes affect lying ascriptions, the possible effect size is likely to be small. Regardless of the stakes being high or low, participants classify them as lies more or less equally. We argue that such an outcome is potentially good news for the definitions of lying because it is unlikely that stakes are able to constitute a significant bias in what we consider to be lies.

Future research should explore the relationship between stakes and deontic status. In LYING, we manipulated either the stakes (high vs low) or the deontic status (forbidden vs obligatory to lie), treating the latter as a kind of moral stakes. However, it is possible to conceive of both low and high stakes scenarios that have the same deontic status. For example, it is generally forbidden to lie to one's parents. One might tell a low-stakes lie by saying, "Jane was here. She is a smoker," to hide one's smoking habit from one's parent. Contrast this with a case in which Jill, after suffering an accident at home, tells the arriving doctor, "Jane was here. She is a smoker." In this scenario, the stakes are intuitively higher: it is extremely important for the doctor to understand the full picture to make an accurate diagnosis. Both cases involve lying in contexts where it is forbidden to lie, yet intuitively differ in stakes. This suggests that LYING may not have explored the full range of possible combinations. Future studies could vary both stakes (low vs high) and deontic status (forbidden vs obligatory to lie).

Besides, as Francis and colleagues emphasize, "stakes or importance are not binary properties—they are scalar: whether a situation is "high" or "low" stakes is a matter of degree." (Francis et al., 2019, p. 427). Thus, future studies may test the influence of stakes across a broader spectrum—for example, by introducing more graded levels of stakes within a single experimental design. This would allow for a more fine-grained analysis of how varying degrees of stakes might (or might not) shape lying attributions.

4. GPT

The folk or ordinary language concept of lying is also the focus of this article. But now the conversations do not only happen between humans: AI has become a conversational agent (Liu et al., 2024). Large Language Models (LLMs), such as GPT-3.5 and its later version, GPT-4, have become indistinguishable from humans in keeping a conversation (Jones & Bergen, 2024). These models, developed by OpenAI, are trained on vast text data to generate human-like responses (*GPT-4 | OpenAI*, n.d.). This article explores how GPT-4 *uses* the concepts of lying and falsity. It does this by comparing the LLM-generated responses with those of humans.

4.1 Motivation

The primary motivation for the research in this paper is the growing concern about the proliferation of mis- and disinformation, particularly as AI-generated content becomes increasingly indistinguishable from human-created content (Brown et al., 2020; Jakesch et al., 2023; Porter & Machery, 2024). Given the scale of this challenge, relying solely on human efforts to mitigate misinformation is impractical and an AI-based solution is needed. But what concept of lying does it possess?

As with LYING, this paper centres on the folk or ordinary language concept of lying (in addition, it also explores the concept of falsity), but it pushes further by contrasting these human concepts with those implicitly operated by GPT-4. LLMs are trained on large amounts of, mostly textual, data (Buyl et al., 2024), and their outputs mirror the statistical patterns embedded in their training sets. Vallor (2024) deploys a mirror metaphor to describe this process, emphasizing that what LLMs produce is a reflection (sometimes a distortion) of existing patterns in human behaviour. It is hard to deny that GPT-4 *uses* the concept of lying in some way, and this paper explores such use.

Given that LLMs can extract and reproduce subtle patterns such as gender bias even when no explicit gender marker is provided (see Dong et al., 2023) there is reason to think they might similarly internalize some salient patterns of lying attributions. Such patterns might have been easily overlooked by philosophers (like the *physiological* use of '(in)voluntarily' was overlooked by Austin and Ryle). Thus, it makes sense to explore what patterns there are within GPT's lying and falsity attributions.

4.2 Description of Experiment

In brief, we made GPT perform the same kind of task as human participants do when they participate in experiments. We investigated the extent to which the concepts of lying and falsity employed by GPT-4 correlate with those of humans.

Quantitative experimental methods have been shown to be successful in capturing relevant characteristics of the folk concept of lying (Wiegmann & Meibauer, 2019). We assumed that an LLM can act as a participant in such an experiment: it can be provided with instructions in the same way human participants are, and it can often do the same tasks (*read* a scenario and *answer* a few questions afterwards). The variability of LLM responses can be addressed using the same statistical tools as are used when running experiments with humans.

For the above-described reasons, the method mimics the human-based experiment as closely as possible. We took (Wiegmann et al., 2024)—a cross-cultural study on lying and falsity—as a basis for comparison. That study employed ten scenarios of deceptive implicatures (plus a control scenario) and was conducted in six languages (English, German, Spanish, Chinese, Japanese, and Hebrew) in 10 countries. Participants evaluated all ten vignettes presented in a random order, with the control scenario being the last. There were two conditions: lying (Did x lie? Yes/No) and falsity (What x said is… True/False); before this crucial question, participants were asked to morally evaluate the protagonist's action from 'very bad' to 'not bad at all' (a 7-point Likert item). The third question was to evaluate the confidence of their responses.

We created a Python script that automatically sends requests to the GPT-4 model and records the LLM's responses. We used the model's API to automate the process. The model was instructed to reply with numbers and to treat the response options as the extremes of the Likert item. For example, the role message contained the following: "You should respond with a number between 1 and 7, considering the middle ground as well" (GPT, p.3). To simulate the between-subject element of the vignettes, the model received the message history containing up to three vignettes and the model's responses to all questions for these vignettes. Due to token limitation, it was impossible to keep the message history of all vignettes because the message to the model was getting too big to be processed. In our experiment, GPT-4 played the same role as ordinary participants do when they provide responses to the scenarios. In total, we gathered 30 such responses for each of the eleven scenarios in six languages and then analyzed the resulting dataset. Then, we compared it to the human data reported by Wiegmann and colleagues and calculated correlations.

4.3 Main Findings and Future Research

Statement for the defence: Strong correlations exist between human and GPT-4 judgements regarding lying and falsity, but there are systematic misalignments.

We found a strong but far from perfect correlation between GPT and Human lying and falsity attributions (see Table 1 in GPT). Overall, the lying attributions rating by GPT 71% is significantly higher than that of human participants 57%. It considers implicatures to be lies more often than humans do. The falsity attributions pattern is mostly reversed (also significant): 42% versus 49%, so less often GPT considers 'what is said' to be false. Besides, GPT constantly provides

higher moral and confidence judgments than humans. If GPT-4 is employed as a tool for detecting misinformation, it would more frequently than ordinary humans classify things as lies and less often than humans when it comes to falsehoods.

The rapid evolution of LLMs requires future evaluation of newer systems. Following GPT-4, OpenAI has already introduced models such as GPT-4.5 or o1; there are also members of Anthropic's Claude series, including Claude 3.7 Sonnet, members of Google's Gemini series, e.g., Gemini 1.5 Pro, and DeepSeek's open-source models, like DeepSeek-R1. Exploring more models, as has already been done by (Almeida et al., 2024), can help us identify the pros and cons of each of the models for addressing the misinformation problem. Besides, our findings might be used to fine-tune existing models, achieving more substantial alignment with human concepts.

5. PEJORATIVE

Austin's notion of *performatives*: utterances that do something rather than merely describe a state of affairs (such as promising, apologizing, or naming) motivated the research of what can be done with words. Many uses of slurs are not merely negative descriptions but actions: they insult and exclude. The last article of the dissertation has explored the ethical profile of an Estonian ethnic pejorative—'tibla'—and attempted to identify contexts when use of the target term is laudable and whether such uses are nonetheless subject to critique.

5.1 Motivation

Speaking about a growing interest in research into expressives (and slurs in particular), Cepollaro (2023) notes:

> This lively debate often relies on a range of empirical claims, grounded in the philosopher's intuitions: "slurs are derogatory in this context (or not)", "epithets have a distinctive power to affect people's beliefs and attitudes", "reporting a slurring utterance backfires", and so on and so forth. Many of these questions can and should be addressed on empirical grounds too. Yet, very few philosophers went for an experimental approach. (Cepollaro, 2023, pp. 173–174)

Cepollaro's (2023) observation echoes the problem of verifying intuitions in OLP. One can easily conceive a Mates-style challenge to the researchers of slurs who back up their claims only by their intuition. It is crucial to distinguish between the users of a slur and its target group. For instance, users may insist that their use of the term is innocuous, while members of the target group may experience it as offensive or exclusionary. Whose intuitions should count when attempting to describe the ordinary meaning of slurs and pejoratives? Often, these groups diverge in their judgments about ordinary use. This disagreement may run even deeper than the one we previously saw between Austin and Ryle.

This article adopted an empirical approach to researching 'tibla'. This word is considered a slur by many of those who fall within its extension (mainly Russian-speaking people living in Estonia), but there are others who consider it a useful word, whilst simultaneously denying that they are prejudiced. We worked from the speculation that the pejorative that appeared in the Soviet time refers to people possessing these characteristics: (i) being ethnic Russian or a Russian speaker, and (ii) having a chauvinist attitude towards Estonia and its culture. While (i) is the ethnic component that constitutes a protected characteristic (no one should be discriminated against based on ethnicity, sex, race, etc., and such discrimination is prosecuted), (ii) validates an assumption of laudable uses: it can regulate anti-social behaviour of targets or pre-emptively contribute into adjustments of their behaviour (e.g., someone will integrate into Estonian society more because they do not want to be called a tibla—similar to how the threat of being called or

thought of as a scrooge or loudmouth might encourage people to be more generous and to keep the secrets they've been entrusted with). These components of the meaning of 'tibla' make the target expression special. It has a hybrid meaning that consists of being similar to both 'ethnic slurs' and 'cognitive-behavioural pejoratives' at the same time.

We aimed to conduct an exploratory case study to describe the ethical profile of 'tibla'. This research task was addressed by providing answers to the following sub-questions:

(1) Are there any plausible examples of laudable uses of the Estonian pejorative when used as an ethnic-cognitive-behavioural pejorative?

(2) Are there critiques to which the Estonian pejorative is susceptible even when used as an ethnic-cognitive-behavioural pejorative?

(3) Are these critiques applicable to the plausibly laudable uses of the Estonian pejorative? (PEJORATIVE, p. 2)

5.2 Description of the Study

The first task of this paper was to identify possibly laudable use(s) of the word 'tibla'. We chose to conduct interviews and talk directly to people who may have witnessed how the pejorative was used in the Soviet time. Qualitative methods appeared to be the best fit for our purpose. The pejorative was censored during the occupation, so we could not draw upon corpora. There was not enough documented data on how the word was used to generate a hypothesis for a quantitative study, plus we considered it very unlikely that we would be able to get a large enough sample for an experiment with high power. Thus, we interviewed people in either Estonian or Russian to get some insights into how the target expression was used. In the semi-structured interviews, participants were asked:

(1) for a gloss on the word's meaning;

(2) if ethnicity was required, was it all that was required to fall into the word's extension;

(3) if cognitive-behavioural traits were required, what these traits were;

(4) for examples of contexts in which the word was used. (PEJORATIVE, p. 4)

The questions attempted to identify criteria for falling within the extension of 'tibla': whether the motivation for calling one 'tibla' was to emphasize one's misbehaviour (like being chauvinistic) or it might have been reduced to merely one's ethnicity. My role was in managing and executing the Russian-speaking part of that project. In the first part, we tried to recruit people who could recall how the pejorative was used in Soviet times. We could not find suitable Russian-speaking interviewees, and the only person I interviewed who was old enough to have memories from this period was not familiar with the target expression. The first part (interviewing older people about the scenarios in which 'tibla' was used)

helped us identify when 'tibla' might have been used for laudable purposes. However, "we were determined to gather accounts from people whose ethnicity could make them potential targets of the word, believing (on broadly standpoint epistemic grounds) that they would be best placed to identify fault in use of the pejorative." (PEJORATIVE, p. 10). So, we conducted more interviews with younger people to identify critiques of the identified uses. The recorded interviews were conducted via Zoom and lasted from 10 to 30 minutes, then transcribed. All identifying information was later removed and coded to keep the identities of interviewees private. The interviews were also translated into English and shared in the project's repository: https://osf.io/km835/. Using this method, we were able to find evidence that even if there are laudable uses of 'tibla' when used as an ethnic-cognitive-behavioural pejorative, even when used in this way, it is still possible to cause harm by using the word. We detail how this is possible. This is an important finding. Everyone agrees that use of the word as an ethnic slur is harmful. But sometimes the presence of the cognitive-behavioural component in a use of the word is presented as if it demonstrated that nothing criticizable could have been done in that use of the word. But this is a mistake. We identify four plausibly criticizable actions that can be performed by use of the word with its hybrid meaning.

5.3 Main Findings and Future Research

Statement for the defence: There are laudable uses of 'tibla' when used as an ethnic-cognitive-behavioural pejorative, bit it is still possible to cause harm by using the word.

In the first stage of this project, we identified a list of cognitive-behavioural traits associated with 'tibla': Chauvinist, Integration Maladaptee, Socially Delinquent, Soviet Mentality, Structurally Empowered. When used between Estonians or muttered to oneself, the pejorative might have been used for "the recognition of an injustice without slipping into another" (PEJORATIVE, p. 9) or for "making injustices visible" (PEJORATIVE, p. 10) to other generations. When used in these scenarios, the use of 'tibla' serves what is arguably a laudable purpose. The second stage identified four potential critiques of use the word with a hybrid meaning: first, even when 'tibla' is used as an ethnic-cognitive-behavioural hybrid pejorative, it might be used as a tool for excluding people who bear some mark of Russian ethnicity where people who lack such marks would not be excluded; secondly, it is often accompanied by remarks implying that improper behaviour *derives from* one's ethnicity (i.e. it is essentialistic); thirdly, it can be used for bullying; finally, even before the pejorative is used, a possessor of markers of Russian ethnicity (e.g., traditionally Russian name) may be subjected to alienating assessment. These critiques are applicable even to the arguably laudable uses of 'tibla'.

The paper, being exploratory in nature, was successful in identifying characteristic features of some uses of 'tibla'. The final section of the paper included a

use scenario that is based on the findings. We argued that a newspaper editorial note justifying the use of 'tibla' in a song used by the Estonian Defence League is unsuccessful as it is mistaken in identifying the extension of the slur (that appears to be broader than the editorial of *Postimees* (a newspaper) assumes). As long as the pejorative is used, there are opportunities for future applications of our findings.

6. CONCLUSION

OL philosophers used to make claims about how ordinary speakers use language (e.g., "we say X but not Y"), but they had no systematic method to verify these claims. They relied on their intuitive "acute ear" for linguistic distinctions, but this provided no way to verify whether their reports of ordinary usage were accurate. In turn, it complicates dealing with disagreements between OL philosophers. The rise of X-Phi has contributed to an OLP revival by adding a range of empirical methods to the philosophical toolbox. Contemporary OLP has found a great ally in X-Phi. The dissertation makes several contributions to the debates in the intersection between contemporary OLP and X-Phi.

The knowledge attribution debates exemplify the disagreement problem that plagued traditional OLP. Multiple experimental designs (evidence-fixed and evidence-seeking) produced conflicting results about whether stakes influence knowledge attributions. The retraction-based experimental design employed in KNOWLEDGE addressed this challenge. Rather than asking participants to evaluate a given statement or protagonist's evidence, the design tested how people actually *use* knowledge attributions: making and retracting knowledge claims. The design consistently found a medium-to-large stakes effect on knowledge attributions. Besides, Experiment 1 successfully verified the findings of Dinges and Zakkou, demonstrating the efficiency of experimental methods in addressing Mates' critique. Given the stakes effect found for knowledge attributions, LYING explored whether a similar effect holds for lying attributions. The folk concept of lying appears to be robust and insensitive to stakes. Traditional OLP relied on how ordinary speakers (humans) use language. The emergence of LLMs creates new possibilities for capturing ordinary use. LLMs operate or *use* relevant concepts themselves and may internalize ordinary language patterns. Treating LLM as a participant in the same experiment as conducted with humans, GPT has identified strong correlations between human and GPT-4 judgments regarding lying and falsity (but also systematic misalignments). Some questions about ordinary language use cannot be addressed through controlled experiments due to practical constraints (censorship, historical usage, sensitive topics, small populations). Semi-structured interviews with purposive sampling allowed us to find laudable uses of the Estonian pejorative 'tibla' in PEJORATIVE. These uses might also be harmful.

Overall, "reports of the death of ordinary language philosophy are greatly exaggerated" (Hansen, 2014, p. 564), and the label of ordinary language philosophy that used to be treated as derogatory can—and, perhaps, should—be proudly reclaimed. With the range of methods demonstrated in this dissertation, OL philosophers have a diverse toolbox for dealing with many questions (knowledge, lying, AI, slurs, etc.), so the future of contemporary OLP seems bright.

REFERENCES

Almeida, G. F. C. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & Araújo, M. de. (2024). Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, *333*. https://doi.org/10.1016/j.artint.2024.104145

Austin, J. L. (1957). A Plea for Excuses: The Presidential Address. *Proceedings of the Aristotelian Society*, *57*(1). https://doi.org/10.1093/aristotelian/57.1.1

Austin, J. L. (1962). *How to Do Things with Words* (J. O. Urmson, Ed.). Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *2020-December*. https://arxiv.org/abs/2005.14165v4

Buckwalter, W. (2010). Knowledge Isn't Closed on Saturday: A Study in Ordinary Language. *Review of Philosophy and Psychology*, *1*(3), 395–406. https://doi.org/10.1007/s13164-010-0030-3

Buckwalter, W., & Schaffer, J. (2015). Knowledge, stakes, and mistakes. *Nous*, *49*(2), 201–234. https://doi.org/10.1111/NOUS.12017

Buyl, M., Rogiers, A., Noels, S., Dominguez-Catena, I., Heiter, E., Romero, R., Johary, I., Mara, A.-C., Lijffijt, J., & De Bie, T. (2024). *Large Language Models Reflect the Ideology of their Creators*. http://arxiv.org/abs/2410.18417

Cappelen, H., & McKeever, M. (2022). In defense of ordinary language philosophy. *Metaphilosophy*, *53*(2–3), 221–237. https://doi.org/10.1111/meta.12548

Carson, T. L. (2010). *Lying and Deception : Theory and Practice*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199577415.001.0001

Cavell, S. (1958). Must we mean what we say? *Inquiry (United Kingdom)*, *1*(1–4). https://doi.org/10.1080/00201745808601279

Cepollaro, B. (2023). Experimentally-Informed Philosophy of Hate Speech. *Logic, Argumentation and Reasoning*, *33*, 173–187. https://doi.org/10.1007/978-3-031-28908-8_9

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2021). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44. https://doi.org/10.1007/S13164-018-0400-9

DeRose, K. (1992). Contextualism and Knowledge Attributions. *Philosophy and Phenomenological Research*, *52*(4), 913. https://doi.org/10.2307/2107917

DeRose, K. (2005). The ordinary language basis for contextualism, and the new invariantism. *Philosophical Quarterly*, *55*(219), 172–198. https://doi.org/10.1111/J.0031-8094.2005.00394.X

DeRose, K. (2009). The Case for Contextualism: Knowledge, Skepticism, and Context. In *The Case for Contextualism: Knowledge, Skepticism, and Context, Vol. 1*. Oxford University Press. https://doi.org/10.1093/ACPROF:OSO/9780199564460.001.0001

Dinges, A., & Zakkou, J. (2021). Much at stake in knowledge. *Mind & Language*, *36*(5), 729–749. https://doi.org/10.1111/MILA.12300

Dong, X., Wang, Y., Yu, P. S., & Caverlee, J. (2023). *Probing Explicit and Implicit Gender Bias through LLM Conditional Text Generation*. https://arxiv.org/pdf/2311.00306

Feltz, A., & Zarpentine, C. (2010). Do you know more when it matters less? *Philosophical Psychology*, *23*(5), 683–706. https://doi.org/10.1080/09515089.2010.514572

Francis, K. B., & Beaman, C. P. (2023). The role of confidence in knowledge ascriptions: an evidence-seeking approach. *Synthese*, *202*(2), 1–15. https://doi.org/10.1007/S11229-023-04236-W

Francis, K. B., Beaman, P., & Hansen, N. (2019). Stakes, Scales, and Skepticism. *Ergo, an Open Access Journal of Philosophy*, *6*(20201214). https://doi.org/10.3998/ergo.12405314.0006.016

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology*, *72*(6), 1429–1439. https://doi.org/10.1037/0022-3514.72.6.1429

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384–394. https://doi.org/10.1257/0002828053828662

Gneezy, U., & Kajackaite, A. (2020). Externalities, stakes, and lying. *Journal of Economic Behavior and Organization*, *178*, 629–643. https://doi.org/10.1016/J.JEBO.2020.08.020

GPT-4 | OpenAI. (n.d.). Retrieved March 5, 2025, from https://openai.com/index/gpt-4-research/

Grindrod, J., Andow, J., & Hansen, N. (2019). Third-person knowledge ascriptions: A crucial experiment for contextualism. *Mind and Language*, *34*(2), 158–182. https://doi.org/10.1111/mila.12196

Gylfason, H. F., Vésteinsdóttir, V., Kristinsson, K., Asgeirsdottir, T. L., & Schram, A. (2023). Gender differences in lying: The role of stakes. *Economics Letters*, *222*. https://doi.org/10.1016/J.ECONLET.2022.110926

Hansen, N. (2012). On an Alleged Truth/Falsity Asymmetry in Context Shifting Experiments. *The Philosophical Quarterly*, *62*(248), 530–545. https://doi.org/10.1111/J.1467-9213.2012.00059.X

Hansen, N. (2014). Contemporary ordinary language philosophy. *Philosophy Compass*, *9*(8), 556–569. https://doi.org/10.1111/phc3.12152

Hansen, N., & Chemla, E. (2013). Experimenting on Contextualism. *Mind & Language*, *28*(3), 286–321. https://doi.org/10.1111/MILA.12019

Hansen, N., & Chemla, E. (2015). Linguistic Experiments and Ordinary Language Philosophy. *Ratio*, *28*(4), 422–445. https://doi.org/10.1111/RATI.12112

Holguín, B. (2021). Lying and knowing. *Synthese*, *198*(6), 5351–5371. https://doi.org/10.1007/s11229-019-02407-2

Horvath, J. (2023). Intuitions in experimental philosophy. *The Compact Compendium of Experimental Philosophy*, 71–100. https://doi.org/10.1515/9783110716931-005

Hu, F., & Ben-Ner, A. (2020). The effects of feedback on lying behavior: Experimental evidence. *Journal of Economic Behavior and Organization*, *171*, 24–34. https://doi.org/10.1016/J.JEBO.2019.12.019

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(11), e2208839120. https://doi.org/10.1073/PNAS.2208839120

Jones, C. R., & Bergen, B. K. (2024). *People cannot distinguish GPT-4 from a human in a Turing test*. https://arxiv.org/pdf/2405.08007v1

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190–194. https://doi.org/10.1111/1467-8284.00419
Knobe, J. (2007). Experimental Philosophy. *Philosophy Compass*, *2*(1), 81–92. https://doi.org/10.1111/j.1747-9991.2006.00050.x
Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–329. https://doi.org/10.1017/S0140525X10000907,
Knobe, J. (2016). Experimental Philosophy Is Cognitive Science. *A Companion to Experimental Philosophy*, 37–52. https://doi.org/10.1002/9781118661666.CH3
Knobe, J., & Nichols, S. (2008). An Experimental Philosophy Manifesto. *Experimental Philosophy*, 3–14. https://doi.org/10.1093/OSO/9780195323252.003.0001
Li, J., & Zhu, X. (2023). Twenty years of experimental philosophy research. *Metaphilosophy*. https://doi.org/10.1111/meta.12602
Liu, N., Chen, L., Tian, X., Zou, W., Chen, K., & Cui, M. (2024). *From LLM to Conversational Agent: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models*. https://arxiv.org/pdf/2401.02777
Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3). https://doi.org/10.1016/j.cognition.2003.10.003
Mates, B. (1958). On the verification of statements about ordinary language. *Inquiry (United Kingdom)*, *1*(1–4), 161–171. https://doi.org/10.1080/00201745808601278
May, J., Sinnott-Armstrong, W., Hull, J. G., & Zimmerman, A. (2010). Practical Interests, Relevant Alternatives, and Knowledge Attributions: An Empirical Study. *Review of Philosophy and Psychology*, *1*(2), 265–273. https://doi.org/10.1007/S13164-009-0014-3
Mizumoto, M. (2023). Experimental Philosophy and Ordinary Language Philosophy. *Logic, Argumentation and Reasoning*, *33*, 31–56. https://doi.org/10.1007/978-3-031-28908-8_3
O'Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, *33*(6), 530–538. https://doi.org/10.1007/S10979-008-9166-4
Pinillos, Á. (2012). Knowledge, Experiments, and Practical Interests. *Knowledge Ascriptions*. https://doi.org/10.1093/ACPROF:OSO/9780199693702.003.0009
Pinillos, Á. (2016). Experiments on contextualism and interest relative invariantism. *A Companion to Experimental Philosophy*, 349–358. https://doi.org/10.1002/9781118661666.CH24
Pinillos, Á., & Simpson, S. (2014). Experimental Evidence in Support of Anti-Intellectualism About Knowledge. In J. Beebe (Ed.), *Advances in Experimental Epistemology* (pp. 9–44). Bloomsbury Academic.
Porter, B., Barr, K., Bencherifa, A., Buckwalter, W., Deguchi, Y., Fabiano, E., Hashimoto, T., Halamova, J., Homan, J., Karasawa, K., Kanovsky, M., Kim, H., Kiper, J., Lee, M., Liu, X., Mitova, V., Bhaya, R., Pantovic, L., Quintanilla, P., … Machery, E. (2024). A puzzle about knowledge ascriptions. *Nous*. https://doi.org/10.1111/NOUS.12515
Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports 2024 14:1*, *14*(1), 1–12. https://doi.org/10.1038/s41598-024-76900-1
Porter, J. D., & Hansen, N. (2023). A quantitative history of ordinary language philosophy. *Synthese*, *201*(6). https://doi.org/10.1007/s11229-023-04187-2

Reins, L. M., Wiegmann, A., Marchenko, O. P., & Schumski, I. (2023). Lying Without Saying Something False? A Cross-Cultural Investigation of the Folk Concept of Lying in Russian and English Speakers. *Review of Philosophy and Psychology*, *14*(2), 735–762. https://doi.org/10.1007/s13164-021-00587-w

Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., Buchtel, E. E., Chatterjee, A., Cheon, H., Cho, I. R., Cohnitz, D., Cova, F., Dranseika, V., Lagos, Á. E., Ghadakpour, L., Grinberg, M., Hannikainen, I., Hashimoto, T., Horowitz, A., … Zhu, J. (2019). Nothing at Stake in Knowledge*. *Nous*, *53*(1), 224–247. https://doi.org/10.1111/NOUS.12211

Ryle, G. (2009). The concept of mind: 60th anniversary edition. In *The Concept of Mind: 60th Anniversary Edition*. Routledge. https://doi.org/10.4324/9780203875858

Saul, J. M. (2012). Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and in Ethics. In *Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and in Ethics*. Oxford University Press. https://doi.org/10.1093/ACPROF:OSO/9780199603688.001.0001

Schaffer, J., & Knobe, J. (2012). Contrastive knowledge surveyed. *Nous*, *46*(4), 675–708. https://doi.org/10.1111/J.1468-0068.2010.00795.X

Sripada, C. S., & Stanley, J. (2012). Empirical tests of interest-relative invariantism. *Episteme*, *9*(1), 3–26. https://doi.org/10.1017/epi.2011.2

Stokke, A. (2013). LYING AND ASSERTING. *The Journal of Philosophy*, *110*(1), 33–60. https://doi.org/10.5840/jphil2013110144

Sytsma, J., Ulatowski, J., & Gonnerman, C. (2023). History and philosophy of experimental philosophy: All in the family. In *The Compact Compendium of Experimental Philosophy* (pp. 9–38). De Gruyter. https://doi.org/10.1515/9783110716931-003

Turri, J. (2016). Knowledge and the norm of assertion: An essay in philosophical science. In *Knowledge and the Norm of Assertion: An Essay in Philosophical Science*. Open Book Publishers. https://doi.org/10.11647/OBP.0083

Turri, J. (2017). Epistemic Contextualism: An Idle Hypothesis. *Australasian Journal of Philosophy*, *95*(1), 141–156. https://doi.org/10.1080/00048402.2016.1153684

Vallor, Shannon. (2024). *The AI mirror : how to reclaim our humanity in an age of machine thinking*. Oxford University Press.

Viebahn, E. (2021). The Lying-Misleading Distinction: A Commitment-Based Approach. *Journal of Philosophy*, *118*(6), 289–319. https://doi.org/10.5840/JPHIL2021118621

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and Epistemic Intuitions. *Philosophical Topics*, *29*(1/2), 429–460. https://doi.org/10.5840/philtopics2001291/217

Weissman, B., & Terkourafi, M. (2019). Are false implicatures lies? An empirical investigation. *Mind & Language*, *34*(2), 221–246. https://doi.org/10.1111/MILA.12212

Wiegmann, A. (2023a). Does lying require objective falsity? *Synthese*, *202*(2). https://doi.org/10.1007/s11229-023-04291-3

Wiegmann, A. (2023b). Lying with deceptive implicatures? Solving a puzzle about conflicting results. *Analysis (United Kingdom)*, *83*(1), 107–118. https://doi.org/10.1093/analys/anac037

Wiegmann, A., & Meibauer, J. (2019). The folk concept of lying. *Philosophy Compass*, *14*(8), e12620. https://doi.org/10.1111/PHC3.12620

Wiegmann, A., Reins, L. M., Mizumoto, M., Erut, A., Li, Q., & Orr, S. (2024). *The role of implicit content for lie and truth judgments: A cross-cultural study*. https://doi.org/10.31234/OSF.IO/GDC2Z

Williamson, T. (2016). Philosophical Criticisms of Experimental Philosophy. *A Companion to Experimental Philosophy*, 22–36. https://doi.org/10.1002/9781118661666.CH2

Wright Whelan, C., Wagstaff, G., & Wheatcroft, J. M. (2015). High stakes lies: police and non-police accuracy in detecting deception. *Psychology, Crime and Law*, *21*(2), 127–138. https://doi.org/10.1080/1068316X.2014.935777

Wu, S. (2023). Are Folks Purists or Pragmatic Encroachers? New Discoveries of Relation between Knowledge and Action from Experimental Philosophy. *Episteme*, 1–29. https://doi.org/10.1017/EPI.2023.56

Zahorec, M., Bishop, R., Hansen, N., Schwenkler, J., & Sytsma, J. (2023). Linguistic Corpora and Ordinary Language: On the Dispute Between Ryle and Austin About the Use of 'Voluntary', 'Involuntary', 'Voluntarily', and 'Involuntarily.' *Logic, Argumentation and Reasoning*, *33*, 121–149. https://doi.org/10.1007/978-3-031-28908-8_7

SUMMARY IN ESTONIAN

Eksperimentaalne panus kaasaegsesse tavakeelefilosoofiasse

Käesolev doktoritöö koondab neli artiklit, mis uurivad erinevaid, kuid omavahel seotud teemasid kaasaegses tavakeele filosoofias, kasutades mitmesuguseid empiirilisi meetodeid. Eksperimentaalse filosoofia (x-phi) tekkimine 2000. aastate alguses on andnud tavakeele filosoofiale metodoloogilised vahendid, et vastata mõningatele selle kriitikale. Need neli artiklit uurivad erinevaid vestluskontekste ja seda, kuidas need erinevad kontekstid mõjutavad meie sõnu kasutamist. Kõik artiklid kasutavad empiirilisi meetodeid, et kirjeldada, kuidas tavalised kõnelejad sihtväljendeid (target expressions) kasutavad. Järgnevalt selgitan iga artikli täpset panust:

Artikkel 1 KNOWLEDGE: „The Stakes Effect: New Evidence from a Retraction-Based Design" uurib, kuidas praktilised panused mõjutavad teadmiste omistamist. Eksperimendid näitavad tugevat „panuste mõju", kus osalejad on märkimisväärselt tõenäolisemad tagasi võtma teadmiste väiteid, kui nad mõistavad, et panused on suured. See kehtib mitte ainult Dinges & Zakkou (2021) eksperimentide puhul (kus tutvustati tagasivõtmisel põhinevat eksperimentaalset disaini), vaid ka otsese replikatsiooni, Banki juhtumi kolmanda isiku versiooni ja eksperimendi puhul, kus rakendati muudatust, mis käsitles mõningaid disaini puudusi. Käesolevas artiklis väidan, et need tulemused seavad kahtluse alla traditsioonilise vaate teadmistele – klassikalise invariantsuse – ja kujutavad endast teatavat väljakutset subjektiivsele invariantsusele. Siiski ei suuda tagasivõtmisel põhinev disain praegusel kujul eristada episteemilise kontekstualismi ja relativismi pakutavaid seletusi.

Artikkel 2 LYING: „The role of stakes in lying: An empirical investigation of the robustness of the folk concept of lying" keskendub valetamise kontseptsioonile, uurides, kas panused mõjutavad inimeste otsuseid selle kohta, kas avaldus klassifitseeritakse valeks. Vähemalt mõned artiklid teadmiste omistamise kontekstisensitiivsusest peetud aruteludes on leidnud panuste mõju. Käesolevas artiklis uuritakse, kas sarnane mõju eksisteerib ka valetamise omistamisel. Pärast nulltulemuse saamist esimeses eksperimendis, mis testib madala ja kõrge panusega stsenaariumide paare, pöörame tähelepanu deontilise staatuse erinevusele: juhtumid, kus valetamine on keelatud või kohustuslik (väidame, et see on õigustatud samm, kuna panused mõistetakse sageli kui peategelase halvemasse olukorda jätmine). Kuigi me täheldasime panuste olulist mõju meie eksperimendis 2, näitas selle kontrollitum versioon – eksperiment 3 – et rahva arusaam valetamisest ei ole oluliselt mõjutatud ei panuste ega deontilise staatuse poolt. See viitab sellele, et rahva arusaam valetamisest on tugev ja ei ole väga tundlik panuste suhtes.

Artikkel 3 GPT: „GPT-4's alignment with human lie and falsity attribution in cases of deceptive implicatures." uurib, nagu pealkiri viitab, seost inimeste valetamise ja vääruse kontseptsioonide ning suure keelemudeli kontseptsioonide

vahel. See kasutab kultuuridevahelist inimeste andmekogu (Wiegmann et al., 2024) võrdlusalusena, et hinnata GPT-4 tulemuslikkust kuues keeles, keskendudes taas petlikele implikatsioonidele. Tulemused näitavad, et kuigi GPT-4 hinnangud korreleeruvad tugevalt inimeste hinnangutega, kalduvad GPT-4 valetamist sagedamini omistama kui inimesed ning väärust harvemini. See uurimus on oluline AI potentsiaalse kasutamise seisukohalt valeinfo tuvastamisel ja parandamisel.

Artikkel 4 PEJORATIVE: „Abhorrent Slurs and Laudable Pejoratives: An Estonian Case Study of the Space Between" on empiiriliselt põhjendatud arutelu pejoratiivse keele eetiliste mõõtmete üle. See esitab juhtumiuuringu eesti pejoratiivse sõna „tibla" kohta, analüüsides selle kasutamist eesti kontekstis. Uuring käsitleb nii pejoratiivse sõna kiiduväärt kasutamise võimalusi kui ka kriitikat, millele see on vastuvõtlik, eriti kui seda kasutatakse etnilise-kognitiivse-käitumusliku pejoratiivina. Artiklis uuritakse, kuidas sellist pejoratiivset sõna saab kasutada olemasolevate ebaõiglusete esiletõstmiseks ja rõhuva võimustruktuuri väljakutseks, jäädes siiski avatuks väärkasutusele (me demonstreerime seda lühikese juhtumiuuringu rakendusega).

Dissertatsioon näitab, et tavakeele filosoofia silt, mida varem peeti halvustavaks, võib (ja võib-olla peaks) uhkelt tagasi võtma. Käesolevas dissertatsioonis tutvustatud meetodite abil on tavakeele filosoofidel mitmekesine tööriistakast tavakontseptsioonidega tegelemiseks, mistõttu kaasaegse tavakeele filosoofia tulevik peab olema helge.

PUBLICATIONS

CURRICULUM VITAE

Name Nikolai Shurakov
Date of birth 20.10.1996
E-mail shurakovnukolay@gmail.com

Education

2020–	PhD in Philosophy University of Tartu, Estonia
2018–2020	Master of Arts in Philosophy University of Tartu, Estonia
2014–2018	Bachelor in Philosophy, with Distinction Saint Petersburg State University, Russia

Professional career

2022– 2025	0.5 Junior Research Fellow University of Tartu, Estonia

A list of Publications

Shurakov N. (forthcoming). The Stakes Effect: New Evidence from a Retraction-Based Experimental Design. Episteme 1–21. https://doi.org/10.1017/epi.2025.10060

Shurakov, N., & Wiegmann, A. (2025). The role of stakes in lying: An empirical investigation of the robustness of the folk concept of lying. Philosophical Psychology.https://doi.org/10.1080/09515089.2025.2506544

Shurakov, N., Kolbe, M., Engelmann, N., & Wiegmann, A. (forthcoming). GPT-4's alignment with human lie and falsity attribution in cases of deceptive implicatures. In V. C. Müller, A. R. Dewey, L. Dung, & G. Löhr (Eds.), Philosophy of Artificial Intelligence: The State of the Art. Synthese Library. Berlin: SpringerNature.

Davies, A., Rahuoja, F. G., & Shurakov, N. (2025). Abhorrent Slurs and Laudable Pejoratives: an Estonian Case Study of the Space Between. Topoi. https://doi.org/10.1007/s11245-025-10163-x

Shurakov, N. (2020) Is a 'free act' a natural kind term? Tomsk State University Journal of Philosophy, Sociology and Political Science №53. pp. 43–50 [In Russian]

Research grants and scholarships

2024	Yale Center for Faith & Culture – Life Worth Living Graduate Fellowship
2023	Deutscher Akademischer Austauschdienst (DAAD)- Short-term Grant (6 months)
2023	Utrecht Network – Young Researcher's Grant

ELULOOKIRJELDUS

Nimi Nikolai Shurakov
Sünniaeg 20.10.1996
E-mail shurakovnukolay@gmail.com

Haridus

2020–	PhD in Philosophy University of Tartu, Estonia
2018–2020	Master of Arts in Philosophy University of Tartu, Estonia
2014–2018	Bachelor in Philosophy, with Distinction Saint Petersburg State University, Russia

Akadeemiline töökogemus

2022–2025	0.5 Junior Research Fellow University of Tartu, Estonia

Publikatsioonid

Shurakov N. (forthcoming). The Stakes Effect: New Evidence from a Retraction-Based Experimental Design. Episteme 1–21. https://doi.org/10.1017/epi.2025.10060

Shurakov, N., ja Wiegmann, A. (2025). The role of stakes in lying: An empirical investigation of the robustness of the folk concept of lying. Philosophical Psychology.https://doi.org/10.1080/09515089.2025.2506544

Shurakov, N., Kolbe, M., Engelmann, N., ja Wiegmann, A. (forthcoming). GPT-4's alignment with human lie and falsity attribution in cases of deceptive implicatures. In V. C. Müller, A. R. Dewey, L. Dung, ja G. Löhr (Eds.), Philosophy of Artificial Intelligence: The State of the Art. Synthese Library. Berlin: SpringerNature.

Davies, A., Rahuoja, F. G., ja Shurakov, N. (2025). Abhorrent Slurs and Laudable Pejoratives: an Estonian Case Study of the Space Between. Topoi. https://doi.org/10.1007/s11245-025-10163-x

Shurakov, N. (2020) Is a 'free act' a natural kind term? Tomsk State University Journal of Philosophy, Sociology and Political Science №53. pp. 43–50 [In Russian]

Uurimistoetused ja -stipendiumid

2024	Yale Center for Faith & Culture – Life Worth Living Graduate Fellowship
2023	Deutscher Akademischer Austauschdienst (DAAD)- Short-term Grant (6 months)
2023	Utrecht Network – Young Researcher's Grant

DISSERTATIONES PHILOSOPHICAE UNIVERSITATIS TARTUENSIS

1. **Jüri Eintalu.** The problem of induction: the presuppositions revisited. Tartu, 2001.
2. **Roomet Jakapi.** Berkeley, mysteries, and meaning: a critique of the non-cognitivist interpretation. Tartu, 2002.
3. **Endla Lõhkivi.** The sociology of scientific knowledge: a philosophical perspective. Tartu, 2002.
4. **Kadri Simm.** Benefit-sharing: an inquiry into justification. Tartu, 2005.
5. **Marek Volt.** The epistemic and logical role of definition in the evaluation of art. Tartu, 2007.
6. **Aive Pevkur.** Professional ethics: philosophy and practice. Tartu, 2011.
7. **Toomas Lott.** Plato on Belief (*doxa*) *Theaetetus* 184B–187A. Tartu, 2012, 208 p.
8. **Jaanus Sooväli.** Decision as Heresy. Tartu, 2013, 153 p.
9. **Ave Mets.** Normativity of scientific laws. Tartu, 2013, 217 p.
10. **Vivian Bohl.** How do we understand others? Beyond theories of mind-reading and interactionism. Tartu, 2014, 184 p.
11. **Uku Tooming.** The Communicative Significance of Beliefs and Desires. Tartu, 2014, 208 p.
12. **Andrus Tool.** Objektiivsuse teema Wilhelm Dilthey vaimuteadustefilosoofias. Tartu, 2014, 329 p.
13. **Janar Mihkelsaar.** Giorgio Agamben and Post-Foundational Political Ontology. Tartu, 2015, 169 p.
14. **Jaana Eigi.** The Social Organisation of Science as a Question for Philosophy of Science. Tartu, 2016, 282 p.
15. **Tarmo Tirol.** Subjektiivsus Jean-Paul Sartre'i eksistentsialismis. Tartu, 2017, 253 p.
16. **Eve Kitsik.** Revisionary Ontology: Improving Concepts to Improve Beliefs. Tartu, 2018, 196 p.
17. **Juhan Hellerma.** Mapping Time: Analysis of Contemporary Theories of Historical Temporality. Tartu, 2020, 138 p.
18. **Henrik Sova.** The Structure of a Consistent Global Pragmatism. Tartu, 2020, 87 p.
19. **Riin Kõiv.** The Content and Implications of Nativist Claims. A Philosophical Analysis. Tartu, 2021, 182 p.
20. **Lisa Zorzato.** Towards a Realist Approach to the Challenge of Fictional Models: Augmented Determination. Tartu, 2023, 150 p.
21. **Jay Zameska.** The Ethics of Public Health: Balancing the Interests of Populations and Individuals. Tartu, 2023, 122 p.
22. **Semen Reshenin.** Individual Self-Determination in Friedrich Schiller's Philosophical Writings. Tartu, 2023, 228 p.

23. **Sergei Sazonov.** The Entrepreneurial Theory of Ownership. Tartu, 2024, 191 p.
24. **Juho Lindholm.** Knowledge and Practice. A Criticism of Epistemology. Tartu, 2024, 372 p.
25. **Heidy Meriste.** The Emotion of Guilt: Integrating Cases Without Perceived Wrongdoing. Tartu, 2025, 199 p.